Перспективные проекты многопрофильных систем совместного доступа к данным, накопленным в различных инфраструктурах мира

(Аналитический обзор)

Л. А. Калиниченко[1], А. А. Вольнова[2], Е. П. Гордов[3], В. Н. Захаров[4], Н. Н. Киселева[5], Д. А. Ковалева[6], О. Ю. Малков[7], И. Г. Окладников[8], А. С. Позаненко[9], Н. В. Пономарева[10], Н. А.Скворцов[11], С. А. Ступников[12], А. З. Фазлиев[13]

1 Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; факультет вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова, leonidandk@gmail.com
2 Институт космических исследований Российской академии наук, alinusss@gmail.com
3 Международный исследовательский центр климатоэкологических исследований Института мониторинга климатических и экологических систем Сибирского отделения Российской академии наук, gordov@scert.ru
4 Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, vzakharov@ipiran.ru
5 Институт металлургии и материаловедения им. А.А. Байкова Российской академии наук, kis@imet.ac.ru
6 Институт астрономии Российской академии наук, dana@inasan.ru
7 Институт астрономии Российской академии наук, malkov@inasan.ru
8 Международный исследовательский центр климатоэкологических исследований Института мониторинга климатических и экологических систем Сибирского отделения Российской академии наук, igor.okladnikov@gmail.com
9 Институт космических исследований Российской академии наук, apozanen@iki.rssi.ru
10 Научный центр неврологии, ponomare@yandex.ru
11 Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, nskv@mail.ru
12 Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sstupnikov@ipiran.ru

13 Центр интегрированных информационных систем Института оптики атмосферы Сибирского отделения Российской академии наук, faz@iao.ru

1. Мотивация работы

Проведение исследований, движимых данными, становится неотъемлемой частью различных областей науки, экономики, бизнеса (далее – областей с интенсивным использованием данных – ОИИД). Данные превращаются в стратегический ресурс практически во всех ОИИД, затрагивающий все сферы деятельности людей и определяющий конкурентоспособность, уровень развития науки, промышленности, здравоохранения, обороноспособности страны. Одной из важнейших проблем сохранения и повышения уровня научных исследований в России является обеспечение возможности эффективного доступа исследовательских организаций России к данным, накапливаемым в мире.

Целью настоящего обзора является анализ существующих и планируемых в мире глобальных инфраструктур обеспечения данными многопрофильных (multidisciplinary) областей исследований с интенсивным использованием данных для последующих решений по созданию инфраструктуры совместного доступа к данным, накопленным за рубежом и в России в таких областях.

В подготовке настоящего обзора принимали участие научные сотрудники из ряда организаций РАН, включая Институт проблем информатики Федерального исследовательского центра «Информатика и управление» РАН (ИПИ ФИЦ ИУ РАН), Институт космических исследований РАН (ИКИ РАН), Институт металлургии и материаловедения РАН (ИМЕТ РАН), Институт мониторинга климатических и экологических систем СО РАН (ИМКЭС СО РАН), Институт астрономии РАН (ИНАСАН РАН), Научный центр неврологии РАМН (НЦН РАМН), Институт оптики атмосферы СО РАН (ИОА СО РАН). Одним из результатов совместной работы групп перечисленных организаций была публикация [1], посвященная анализу проблемы доступа к данным в исследованиях с интенсивным использованием данных в России. Настоящая работа поддержана грантом РФФИ 16-07-01028.

2. Открытая наука и планы ее организации в Европе

Наука изменяется под действием стремительного развития информационных и коммуникационных технологий, которые за удивительно короткий промежуток времени произвели в научных методах настоящую революцию. Сейчас научно-исследовательские инфраструктуры предлагают учёным беспрецедентные возможности доступа к источникам данных, уникальным установкам, движимых интенсивным использованием данных, а также постоянно улучшающимся средствам анализа и симуляции. Исследовательские сервисы, процессы и результаты становятся доступными для всех слоёв общества. Генерируются громадные объёмы данных, принося новые необычные возможности инновационного повторного использования их в новых научных, коммерческих контекстах, включая также контексты гражданской науки. Это – Открытая наука. Открытая наука — это ключевой движитель не только научного прогресса, но и экономических и социальных инноваций. Основные особенности Открытой науки можно кратко охарактеризовать так:

- Открытая наука ставит своей целью преобразовать научный процесс с помощью ICT инструментов, сетей и средств связи, чтобы исследования стали более открытыми, глобальными, совместными, творческими и приближенными к обществу.
- Открытая наука занимается вопросами, как исследования проводятся, распространяются, организуются и трансформируются цифровыми методами, сетями и средствами связи. Она полагается на объединённый эффект технологического развития и культурных перемен в сторону коллаборации и открытости исследований.
- Открытая наука делает научные процессы более эффективными, прозрачными и эффективными вводя новые инструменты для научной коллаборации, экспериментов и анализа и делая научные знания более доступными.
- Открытая наука — это ключевой шаг в направлении нового пути открытий, совместного использования и сохранения Знаний.

Три хорошо известных ключевых компонента в развитии Открытой науки включают:

1) Открытый доступ к исследовательским публикациям, обеспечивающий прямое их использование, без каких-либо ограничений, регистраций или подписок.

2) Улучшенное управление исследовательскими данными, покрывающее полный цикл обработки данных, от планирования, сбора и курирования данных до публикации, включения их в анализ и сохранение.

3) Улучшенные е-инфраструктуры позволяют обрабатывать большие коллекции данных, извлекать информацию из научных баз данных и литературы, а также обеспечивает распределённое взаимодействие исследователей на всех уровнях, включая вклад от гражданской науки.

В рамках рамочной программы ЕвроСоюза HORIZON 2020 с 2017 г. Еврокомиссией введено требование, чтобы доступ к публикациям всех проектов этой программы стал открытым, а данные также стали открытыми и доступными. В июле 2016 г. Еврокомиссия опубликовала руководство [2]. Этот документ связан с пилотным проектом Open Research Data Pilot [3], целью которого является улучшение и максимизация доступа и повторного использования исследовательских данных, генерируемых в проектах HORIZON 2020, а также установление баланса между открытостью и защитой научной информации, коммерциализацией и правами интеллектуальной собственности, соображениями приватности, безопасности, наряду с изучением вопросов управления данными и их сохранения.

Другим проектом, связанным с повышением эффективности повторного использования данных, образуемых в рамках проектов программы HORIZON 2020, является the OpenAIRE 2020 Project, поддерживаемый Еврокомиссией. В этом проекте, выполняемом 50 организациями Евросоюза и за его пределами, развивается крупномасштабная инициатива продвижения Открытой науки в направлении повышения эффективности обнаружения и повторного использования публикаций и данных – результатов исследований. Результатом проекта станут потоки работ и сервисы над ценным контентом, содержащим результаты исследований, поддерживающие интероперабельную сеть репозиториев на основании общих руководств.

3. Сводка результатов анализа ситуации в России после 2020 года

Целью исследования [1] явился анализ глобальных тенденций создания массивных коллекций данных в мире и обеспечения возможности совместного использования таких коллекций при решении задач исследования и принятия решений в различных ОИИД в России. Конкретный набор ОИИД, отобранных для анализа, включал астрономию, материаловедение, науки о Земле, геномику и протеомику, нейронауку. По каждой из рассмотренных ОИИД представлены крупные стратегические инициативы США и Европейского Союза (ЕС), примеры крупных коллекций данных в мире до 2025г., известные проекты информационных и телекоммуникационных инфраструктур и центров данных, в том числе:

- Астрономия (большой обзорный телескоп LSST, массив квадратного километра (Square Kilometer Array, SKA) — наиболее амбициозный проект в радиоастрономии, планируемый ESO к пуску в 2024 г. гигантский телескоп E-ELT (с диаметром главного зеркала 39 м), космическая обсерватория Гайя, глубоководный нейтринный телескоп KM3NeT, детекторы гравитационных волн, а также публичные коллекции данных, которые используются в России и требуют улучшения доступа к данным).
- Материаловедение (данные, получаемые в рамках Стратегической инициативы США по Геному Материалов (MGI), коллекции данных в Европе (например, STN, Springer Materials), в США (данные NIST), в Японии (данные в Национальном институте материаловедения)).
- Геномика и протеомика (данные, получаемые при помощи 7400 высокопроизводительных геномных секвенаторов в мире (из них в России находится всего 14 секвенаторов), коллекции геномных данных (данные накапливаются в таких проектах как «1001 геном», «Геном 10К», «Человеческий микробиом», «Атлас генома рака»), атлас протеом человека, развиваемый в Швеции, интегрированное хранилище данных протеомики в Европе, данные проекта

Европейской молекулярной биологической обсерватории ELIXIR, данные Европейского проекта BD2K (от больших биомедицинских данных к знаниям).

- Нейронаука (данные стратегических проектов исследования мозга человека – Европейского (HBP) и проекта США (BRAIN), данные коннектома мозга человека (проект HCP), атласы мозга, совмещающие геномику и нейроанатомию).
- Науки о Земле (данные, получаемые в рамках большого числа проектов в разных странах, в том числе в рамках космических программ в Европе (Copernicus, 30 спутников), в США (программы EOS, EOSDIS), глобальной системы GEOSS, проекта DataONE в США для поддержки совместного использования данных, накапливаемых в многочисленных репозиториях в федеральных сетях. и ряда других проектов, в том числе специализирующихся в области климата Земли.

Всюду, где в этом перечне упоминаются «данные», следует читать «данные и сервисы». В [1] приведены также примеры коллекций данных в названных областях, накопленных в России. В результате анализа планов по созданию крупных коллекций данных в мире в [1] сделаны следующие выводы.

Для достижения эффективного доступа исследовательских организаций России к данным, накапливаемым в мире, с целью их совместного использования с российскими данными в исследовательских проектах России, представляется целесообразной организация целевой междисциплинарной программы для реализации пилотного проекта распределенной инфраструктуры для накопления и анализа данных, совместимой с зарубежными открытыми инфраструктурами в науке.

Одной из первостепенных задач такой программы является анализ и выбор вариантов инфраструктур и платформ для поддержки решения задач анализа больших данных в различных ОИИД, а также для обеспечения доступа исследователей к разнообразным видам данных в мире и совместного междисциплинарного их использования. Настоящий проект является шагом в этом направлении.

Включенный в [1] набор массивных коллекций данных, планируемых к получению в мире, предлагается использовать в качестве ориентира при планировании и развитии исследовательских инфраструктур для накопления и анализа данных, совместимых с зарубежными открытыми инфраструктурами в науке. Рассматриваемые в [1] коллекции данных, цели их создания и научные исследования, планируемые к осуществлению с их помощью, позволяют перейти к постановке задач для решения на базе перспективных информационных и телекоммуникационных инфраструктур, обеспечивающих поддержку открытой науки.

4. Европейские проекты исследовательских инфраструктур и вех

Исследовательские инфраструктуры (Research Infrastructures), создаваемые в ЕС, представляют собой средства, ресурсы или сервисы уникальной природы, которые были идентифицированы в различных областях сообществами исследователей Европы для поддержки соответствующей деятельности на уровне Евросоюза. Подобное определение исследовательской инфраструктуры, включая ассоциированные с ней людские ресурсы, охватывает крупное оборудование или наборы инструментов вместе с содержащими знания ресурсами, такими как коллекции данных, архивы или банки данных. ЦЕРН – родоначальник идеи исследовательских инфраструктур.

Е-инфраструктуры определяются в терминах объединения сетей, гридов, центров данных и сред взаимодействия, намерения по отношению к ним предполагают включение в них центров поддержки операций, регистров сервисов, сервисов достоверных рекомендаций, авторизации сертификатов, тренировочных и консультационных сервисов. Примерами действующих в Европе е-инфраструктур являются GEANT (высокоскоростные сети - http://www.geant.org/), EGI (облачные и грид-вычисления - https://www.egi.eu/), PRACE (суперкомпьютерные вычисления - http://www.prace-ri.eu/), EUDAT (общие сервисы обмена данными в гетерогенных сетях - https://www.eudat.eu/), IDGF (вычисления в толпе - http://idgf-sp.eu/), Open Aire (хранилище научных статей - https://www.openaire.eu/).

Европейский стратегический форум исследовательских инфраструктур (European Strategy Forum on Research Infrastructures, ESFRI) является стратегическим механизмом, образованным в 2002 г. странами — членами ЕС и Еврокомиссией, чтобы способствовать научной интеграции Европы и усилению ее международного влияния. Члены ESFRI назначаются министрами науки стран — членов или ассоциированных членов ЕС, а также включают представителей Еврокомиссии. Они работают совместно для формирования объединенного видения и общей стратегии, включающих в качестве инструментов планирования и реализации новых панъевропейских исследовательских инфраструктур регулярно обновляемые дорожные карты, отчеты и критерии. Подобный стратегический подход нацелен на обеспечение Европы наиболее современными исследовательскими инфраструктурами, отвечающими нуждам быстро развивающихся областей науки, продвижение основанных на знаниях технологий и расширение их применений.

Некоторые инфраструктуры квалифицируются ESFRI как *вехи* (Landmarks) если они достигли фазы реализации и могут быть позиционированы как панъевропейские хабы научного совершенства, генерирующие новые идеи и раздвигающие границы науки и технологии.

Дорожная карта ESFRI 2016 г. [4] включает 21 проект исследовательских инфраструктур и 29 инфраструктур, квалифицированных как вехи. По областям применения они классифицированы так:

- Энергия (4 проекта и 1 веха)
- Окружающая среда (5 проектов и 5 вех)
- Здоровье и пища (8 проектов и 6 вех)
- Физические науки и инженерия (3 проекта и 11 вех)
- Социальные и культурные инновации (1 проект и 5 вех)

Примеры проектов и вех:

- энергетика (проект: EU-SOLARIS – Европейская исследовательская инфраструктура SOLAR для концентрированной солнечной энергии, завершение в 2020 г., 120 миллионов евро; веха: JHR – реактор Юлия Горовица, завершение в 2020 г., 1 миллиард евро);
- окружающая среда (проект: SIOS (Svalbard Integrated Arctic Earth Observation System) – интегрированная система наблюдений Арктики на Шпицбергене, завершение 2020 г., 80 миллионов евро; веха: LifeWatch – исследование биоразнообразия и экосистемы, завершение в 2016 г., 66 миллионов евро);
- здоровье и пища (проект: AnaEE – инфраструктура для анализа и экспериментов в экосистемах, завершение 2018 г., 200 миллионов евро; веха: ELIXIR – распределенная

инфраструктура поддержки информации в науках о жизни, завершение в 2014 г., 125 миллионов евро);

- физические науки (проект: CTA – массив телескопов Черенкова, завершение 2023 г., 297 миллионов евро; веха: E-ELT – Европейский гигантский телескоп, завершение 2024 г., 1 миллиард евро);
- инновации в области социальной сферы и культуры (проект: E-RIHS – Европейская исследовательская инфраструктура в области науки наследия, завершение в 2022 г., 4 миллион евро; веха: DARIAH ERIC – цифровая исследовательская инфраструктура в области искусства и гуманитарной сферы, завершение в 2019 г., 4.3 миллиона евро).

5. Проекты инфраструктур поддержки Открытой науки в Европе

В настоящем обзоре особое внимание уделено перспективным проектам инфраструктур, которые уже организованы или будут организованы в рамках конкурсов HORIZON 2020, нацеленных на организацию открытых исследовательских данных, развитие наук с интенсивным использованием данных, создание крупных инновационных распределенных инфраструктур для совместного использования данных исследователями в разных странах Европы.

Оставаясь в рамках рабочей программы HORIZON 2020 [5] для проведения конкурсов в 2016 – 2017 годах в области исследовательских инфраструктур, включая е-инфраструктуры, в настоящей работе главным образом анализ будет сфокусирован на теме *«eInfrastructures and European Open Science Cloud»* (https://ec.europa.eu/digital-single-market/events/cf/ict-proposers-day-2016/item-display.cfm?id=18473).

В аспекте е-инфраструктур организуются два конкурса [6] по темам:

- EINFRA-12-2017: е-инфраструктуры данных и распределенных вычислений для Открытой науки и
- EINFRA-21-2017: инновации е-инфраструктур, движимых платформами.

EINFRA-12-2017 покрывает две взаимодополняющих области е-инфраструктур, тесно связанных с целью сделать исследовательские данные *обнаруживаемыми, доступными, оцениваемыми, понятными, пригодными для использования и, где возможно, интероперабельными:*

(а) *Защищенные и динамичные данные и е-инфраструктуры для распределённых вычислений*. Вызов состоит в том, чтобы интегрировать на общеевропейском уровне географически и дисциплинарно-распределённые ресурсы ради масштабируемой экономии и достижения эффективности в предоставлении наилучших данных, вычислительных возможностей и сервисов исследовательскому и образовательному сообществу. Эти намерения связаны с программой INFRADEV-04-2016, “*European Open Science Cloud for Research*”.

(б) *Платформы доступа к научной информации и ее сохранения*: поддержка интеграции и консолидации е-инфраструктур для надёжного и постоянного открытого доступа к цифровой научной информации, основанного на существующих инициативных проектах из конца в конец чрез всю Европу (ведомственные и тематические хранилища, агрегаторы и т.д.).

Основная цель инициативы EINFRA-21-2017 состоит в поддержке публичного распространения инновационных HPC-систем, которые должны обеспечить пропускные способности, требуемые в будущих поколениях е-инфраструктур, чтобы отвечать долгосрочным

нуждам исследовательских и образовательных сообществ на отрезке времени 5-10 лет. Ключевыми моментами здесь являются улучшение координации спроса и предложения в европейских HPC экосистемах и поддержка эволюции сервисов е-инфраструктур, основанных на ресурсах данных экзамасштаба.

Оба конкурса EINFRA представляют собой конкретные шаги в реализации проекта European Open Science Cloud [7], предусмотренного в рамках European Cloud Initiative [8], стартовавшей в апреле 2016 года. Проект имеет целью интеграцию и консолидацию е-инфраструктур, образование федерации существующих исследовательских инфраструктур и научных облачных систем, развитие облачных сервисов для Open Science [9]. European Cloud Initiative является частью пакета мер для Digitising European industry, направленных на усиление позиции Европы в инновациях, движимых данными, улучшение их конкурентоспособности и их тенденции к консолидации усилий, а также способствует созданию в Европе Digital Single Market.

5.1 Проект Обменных Сервисов Открытой науки

Существующие сообщества EGI и EUDAT объединяют усилия, чтобы разрешать вопросы текущей фрагментации данных и ландшафта вычислительных е-инфраструктур, и ищут сотрудничества с частными и общественными партнерами с целью создания и расширения тематических сервисов, поддерживающих исследовательские потоки работ Открытой науки. Эти сервисы будут технологически и операционно-интегрированы с обобщенными функциями вычислений, хранения, управления данными и безопасности с целью обеспечения более богатого набора цифровых инструментов для европейских и международных исследовательских сообществ. Управление сервисами, обучение и техническая поддержка также подразумеваются в проекте.

Консорциум EGI & EUDAT готовит предложение, которое будет представлено как часть конкурса H2020 EINFRA-12-2017 (подтема (a), безопасные и динамичные данные и е-инфраструктуры для распределённых вычислений). Целью проекта является неразделимость инфраструктурных сервисов, формирования сообщества пользователей и тренинга. Это может привнести в существующие научные сообщества новые мощные возможности выполнения исследований, а вокруг инструментов и данных могут консолидироваться новые сообщества. Проект направлен на создание истинно открытых и способствующих взаимодействию платформ, пригодных для управления, анализа, совместного и повторного использования, а также сохранения исследовательских данных, на основе которых могут развиваться инновационные сервисы, способные внести новый вклад в мировое научное сообщество.

В следующих подразделах рассматриваются положения о функциональности и начальном состоянии составляющих это предложение инициатив EGI и EUDAT. Для EUDAT также будут представлены оценки ожидаемых результатов развития проекта EUDAT2020, выполняемого с 2015 года.

5.1.1. Федерация EGI

EGI [10] является федерацией центров данных, направленных на предоставление исследователям передовых вычислительных услуг. Она управляется Фондом и Советом EGI. Грид распределённых вычислений был изначально задуман в 1999 году с целью анализа экспериментальных данных, которые должны были поступать с Большого адронного коллайдера в ЦЕРНе. Европейский проект DataGrid, который стартовал в 2001, развивал исследования и разработку грид технологий и продемонстрировал успешность применения гридов в различных областях исследований — физике высоких энергий, наблюдениях Земли и биоинформатике.

Начиная с марта 2004 года, дальнейшая работа по развитию гридов выполнялась серикй проектов EGEE (Enabling Grid for E-sciencE).

С сентября 2007 по декабрь 2009 существовала Европейская инициативная студия дизайна гридов. Фонд EGI, названный EGI.eu, был основан 8 февраля 2010 года с целью координации и поддержки Европейской инфраструктуры гридов — долгосрочного общеевропейского проекта, созданного для поддержки европейских исследовательских сообществ и их международного взаимодействия.

Проект EGI-InSPIRE, нацеленный на создание бесшовной системы, готовой служить требованиям научной деятельности настоящего и будущего, поддерживал функционирование EGI в течение четырех лет до декабря 2014.

Проект EGI-Engage стартовал в марте 2015 для ускорения реализации Open Science Commons путем расширения возможностей европейской магистрали федеративных сервисов для вычислений, хранения данных, коммуникаций, обмена знаниями и опытом, дополняя специфичные для сообществ возможности.

В 2016 году европейская GRID-инфраструктура стала называться EGI. EGI является федерацией провайдеров хранения данных и проведения вычислений, объединенных миссией по поддержке научных исследований и разработок.

Федеративное Облако EGI является облаком IaaS-типа (Infrastructure-as-a-Service – инфраструктура как услуга), состоящим из приватных академических облаков и виртуализированных ресурсов, построенных на основе открытых стандартов. Результатом является новый тип исследовательской е-инфраструктуры, основанной на зрелых сервисах федеративных операций, что делает EGI надежным ресурсом для науки. EGI поддерживает следующие облачные сервисы:

- Cloud Compute – выполнение виртуальных машин с полным контролем над вычислительными ресурсами; возможность выбора предварительно сконфигурированных виртуальных устройств (например, процессор, память, диск, операционная система или программное обеспечение) из каталога, тиражированного через всех поставщиков облачной инфраструктуры EGI;
- Cloud Container Compute – выполнение Docker-контейнеров в легковесной среде виртуализации;
- Training infrastructure – выделенные вычислительные средства и средства хранения для тренинга и обучения.

Федеративная е-инфраструктура EGI финансируется публично и включает (по состоянию на сентябрь 2016):

- 826 500 ядер, доступных для высокопроизводительных вычислений;
- 6 600 ядер, доступных для облачных вычислений;
- 285 PB для оперативного хранения;
- 280 PB для хранения архивов.

Эта инфраструктура включает в себя также федеративные облачные провайдеры EGI и федеративные центры данных EGI.

5.1.2. EUDAT2020

Европейская комиссия поддерживает развитие панъевропейской междисциплинарной инфраструктуры данных в рамках программы Horizon 2020, следуя нескольким ведущим принципам.

Федерализация. Предполагается, что основные действия над данными реализуются в федерациях данных. Они являются сетями репозиториев и центров данных, которые предоставляют структуры для обработки данных и действуют на основе соглашений о легальных или этических правилах, интерфейсах и спецификациях протоколов, а также стека общих сервисов манипулирования данными. Такие центры могут являться членами многих федераций. Координированный подход предполагает, что каждый центр создает описание своих возможностей, а каждая федерация может использовать одни и те же описания для извлечения необходимой информации. Такой подход способствует открытому представлению исследовательских данных и помогает изменять существующую культуру исследований для поддержки совместного использования данных.

Открытое совместное использование данных. Поскольку научные дисциплины интернациональны по своей природе, то критичным является следование международным подходам к снижению барьеров при обмене данными или при их повторном использовании. На этом пути основными препятствиями являются неоднородность данных и языков запросов, способность к пониманию и обнаружению данных, перемещение данных сквозь семантические границы между многозначными контекстами, а также проблемы рассогласования данных (относительно качества, неполноты, абстракции данных).

Европейская инфраструктура данных EUDAT [11] является начальным шагом в этих направлениях. EUDAT объединяет 25 европейских партнеров, включающих центры данных, провайдеры технологий, сообщества исследователей и фондовые агентства из 15 стран. EUDAT предлагает общие сервисы данных в рамках географически распределенной сети, связывающей центры данных и специализированные репозитории, а также решения для поиска, совместного использования, хранения, репликации, стадийности первичных и вторичных данных исследований и выполнения их анализа. Такая сеть образует Совместную инфраструктуру данных (Collaborative Data Infrastructure), обозначаемую далее СИД, которая развивается как сервис-ориентированная, междисциплинарная и устойчивая инфраструктура. Ее сервисы были разработаны в тесном сотрудничестве с более чем 50 междисциплинарными научными сообществами, вовлеченными во все этапы процесса проектирования. Учреждение СИД EUDAT является своевременным в свете предстоящей реализации Европейского облака Открытой науки (European Open Science Cloud), которое нацелено на предоставление открытых и бесшовных услуг для хранения данных, управления, анализа и повторного использования данных исследований в разных научных областях.

Сервисы EUDAT для научных сообществ:

- B2DROP Sync and Exchange Research Data (Синхронизация и обмен исследовательскими данными): персональное облачное решение, основанное на доверенном домене EUDAT CDI, для хранения и совместного использования наборов данных на ранних этапах жизненного цикла исследовательских данных;
- B2SHARE Store and Share Research Data (Хранение и совместное использование исследовательских данных): удобный в пользовании, надежный и защищенный сервис для

научных сообществ, предназначенный для хранения и совместного использования небольших наборов исследовательских данных, полученных из разнообразных источников;

- B2FIND Find Research Data (Поиск исследовательских данных): простой, удобный в пользовании портал для нахождения коллекций исследовательских данных, сохраненных в центрах данных EUDAT и других репозиториях данных;
- B2SAFE Replicate Research Data Safely (Безопасная репликация исследовательских данных): устойчивый, безопасный и высоконадежный сервис для управления данными и их репликацией позволяет репозиториям сообществ и департаментов тиражировать и хранить исследовательские данные на узлах данных EUDAT;
- B2STAGE Get Data to Computation (Извлечение данных для вычислений): надежный, эффективный и простой в использовании сервис для перемещения больших объемов исследовательских данных между узлами данных EUDAT и рабочими областями высокопроизводительных вычислительных систем.

Важным примером консолидированной архитектуры для взаимодействия с инфраструктурами в науках о жизни, позволяющей структурным биологам извлечь пользу из универсальных сервисов, разработанных EUDAT и EGI, является проект West-Life [13], представляющий собой виртуальную исследовательскую среду H2020, которая предоставит сервисы прикладного уровня, приспособленные для сценариев использования в структурной биологии, покрывая все методики эксперимента (например, рентгеноскопия (Xray), электронная криоскопия (cryo-EM), ядерная магниторезонансная томография (NMR), малоугловая рентгеноскопия (SAXS)).

3-го октября 2016 года шестнадцать крупных европейских исследовательских организаций, вычислительных центров и центров данных подписали соглашение о поддержке EUDAT как панъевропейской коллаборативной инфраструктуры данных в течение следующих 10 лет. Организации сплотились для реализации долгосрочного плана устойчивого развития и внесения вклада в разработку, поддержку и развертывание панъевропейских сервисов для исследовательских данных и координацию практик управления исследовательскими данными по всем центрам.

EUDAT2020 [12] — трехлетний большой проект развития СИД, начатый в 2015 г., целями которого являются: поддержка политики Европейской комиссии открытого доступа к данным исследований, достижение интероперабельности существующих в Европе инфраструктур научных исследований (ИНИ) для доступа ученых к сетевым, вычислительным ресурсам и ресурсам данных в различных ИНИ, включая гриды и облачные инфраструктуры. Так, например, будут достигнуты возможности подключения данных в СИД к высокопроизводительным ресурсам, организуемым в рамках PRACE (Partnership for Advanced Computing in Europe), для их анализа или в качестве входных данных моделей и репликации полученных результатов в систему хранения EUDAT; подключения данных в СИД к гридам и облачным ресурсам, поддерживаемым EGI (European Grid Infrastructure); а также федерализации данных при их подключении к ряду европейских инициатив(таких как Nebula, GEANT, TERENA, OpenAIR и др.). При организации EUDAT2020 достигнута договоренность о партнерстве с NDS по образованию совместных пилотных проектов (междисциплинарных и межконтинентальных). В СИД будет поддерживаться функция долгосрочного архивирования данных, репликации, каталогизации, цитируемости данных наряду с обеспечением обнаружения, доступа, повторного использования коллекций и отдельных объектов данных. Функции анализа данных будут поддерживаться ресурсами EGI и PRACE, а также средствами, образуемыми на основе виртуализации вычислительного оборудования центров данных и кластерных платформ.

Специальная программа в рамках EUDAT2020 ориентирована на создание средств оценки качества данных и сертификации репозиториев данных в СИД. EUDAT2020 развивает мультидисциплинарный подход, охватывая сообщества исследователей в гуманитарных областях и в социальных сетях (CLARIN — Common Language Resources and Technology Infractructure, DARIAH, CESSDA), в науках о Земле и атмосфере (EPOS — European Plate Observing System, ICOS, EMSO, VERCE, IAGOS, DRIHM), науке о климате (ENES — European Network for Earth System), биоразнообразии (LifeWatch, LTER, iMarine), науке о жизни (VPH, ELIXIR, BBMRI, ECRIN, INCF, DiXA) и физике (EISCAT, EURO-VO, ISIS, WLCG, PaNdata). Значительное внимание в проекте будет уделено динамическим данным и научным потокам работ, созданию сервисов управления динамическими данными, оставаясь в рамках СИД. Эти исследования будут опираться на сценарии динамического использования данных из ENES и EPOS и обобщения их для анализа будущих динамических данных при решении реальных научных задач. Одним из планируемых результатов будет создание модели и языка представления жизненного цикла данных, сервисных инфраструктур и происхождения данных. Одновременно будут происходить исследования инфраструктурных операций более эффективных, надежных, устойчивых и близких к потребностям научных сообществ. Примерами планируемых задач являются следующие: оценка объектно-ориентированной среды хранения для машин баз данных, центров данных при создании масштабируемой и интероперабельной СИД на основе облачных решений; расширение возможностей уровня долговременного хранения путем применения распределенной графовой базы данных для поддержки отношений между объектами данных вместо собственной базы метаданных, используемой в настоящее время в B2SHARE-сервисе (по замыслу это должно сблизить подходы в СИД с применениями семантического веба, поддержкой происхождения данных и семантического аннотирования).

5.1.3 Особенности планируемого проекта EGI & EUDAT

EGI & EUDAT объединяют усилия для решения проблемы текущей фрагментации данных и ландшафта вычислительных е-инфраструктур и ищут частных и общедоступных партнеров, разрабатывающих и/или предоставляющих тематические услуги, которые поддерживают исследовательские потоки работ Открытой науки. Эти сервисы будут технически интегрированы с общими средствами вычислений, хранения данных, управления данными и безопасности для обеспечения более богатого набора цифровых возможностей для европейских и международных коллаборативных исследований. Управление сервисами, тренинг и техническая поддержка будут также частью этого проекта [14].

Эти партнеры должны быть экспертами в предоставлении сервисов, инструментов или платформ, включая анализ данных, научные приложения, наборы данных, публикации и другие объекты исследования, чтобы эти сервисы можно было легко найти, получить доступ и использовать их снова и снова для получения новых знаний и стимулирования инноваций.

Насколько можно судить, проект будет нацелен на соединение инфраструктурных решений EGI (облачная инфраструктура, центры данных, вычислительные сервисы) с сервисами данных EUDAT, чтобы обеспечить целостное решение для доступа сообщества исследователей к исследовательским данным.

5.2 Инициатива Евросоюза по созданию Европейского облака открытой науки (EOSC)

В этом разделе рассматриваются основные этапы развития программы создания EOSC и связанных с ними публикаций:

- фиксация принципов FAIR данных, ориентированных на поддержку ряда международных инициатив (включающих прежде всего EOSC [8] и проект NIH Big Data to Knowledge (BD2K, http://datascience.nih.gov/bd2k)) и служащих в качестве руководства по приданию всем данным и связанным с ними сервисам свойства быть обнаруживаемыми (Findable), доступными (Accessible), интероперабельными (Interoperable) и повторно используемыми (Reusable) в Интернете данных не только для людей, но, что особенно важно, также и для машин. Принципы FAIR опубликованы в марте 2016 г. [15] и ранее в [16];
- объявление Еврокомиссии в апреле 2016 г. об инициативе Европейского облака [8], которое должно обеспечить Европе глобальное лидерство в экономике, движимой данными;
- заявка на реализацию международного проекта создания Интернета FAIR данных и сервисов, координируемого Нидерландами (июнь 2016 г.);
- первый отчет Экспертной группы высокого уровня, учрежденной Европейской Комиссией, содержащий рекомендации Еврокомиссии по предварительному этапу создания EOSC (октябрь 2016 г.).

Этот перечень завершается разделом, посвященным пилотному проекту исследования возможности развития инфраструктуры EUDAT для обеспечения интероперабельности данных на основе принципов FAIR.

Принципы FAIR были приняты в качестве основополагающих критериев для групп, работающих над созданием исследовательских инфраструктур, политических групп, фондирующих организаций и др., включающих, например, the G20 Hangzhou Concensus, Amsterdam Call for Action on Open Science, the NIH Data Commons и the European Open Science Cloud. Так, поддержка идеи FAIR на саммите G20 в Ханчжоу 4-5 сентября 2016 г. выражена так: «We support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR) principles».

5.2.1 Принципы представления FAIR данных

На протяжении последних двух-трех лет в результате многочисленных обсуждений выкристаллизовалась насущная потребность в создании инфраструктуры, которая поддерживала бы свойства FAIR данных, являющихся результатом исследований (в том числе, в рамках Открытой науки). Участники таких обсуждений из разнообразных областей, представляющих академическое сообщество, промышленность, фондирующие агенства, научные издательства пришли к соглашению о выработке лаконичного и формализуемого набора принципов, объявленного как FAIR Data Principles. В отличие от инициатив, сфокусированных на предоставлении возможностей повторного использования данных людьми, принципы FAIR нацелены на повышение способностей машин автоматически находить и использовать данные в дополнение к поддержке повторного использования данных людьми. Более того, инструменты и потоки работ, связанные с такими данными, также должны обладать свойствами FAIR (быть обнаруживаемыми (Findable), доступными (Accessible), интероперабельными (Interoperable) и повторно используемыми (Reusable)) [15][16]. Данные принципы применимы в общем случае не только к самим данным и сервисам, но и к методам, программам, статьям и документам, относящимся к исследованиям.

К этим четырем был впоследствии добавлен принцип цитируемости данных и других исследовательских объектов. Помимо глобальной идентификации и авторов, цитирование должно содержать информацию о версиях, временных интервалах, используемых фрагментах данных, происхождении используемых вторичных данных. Помимо этого, необходимо понимать, что цитирование может производиться и между сообществами разных предметных областей и должно

учитывать специфику этих сообществ. Таким образом, становится возможным долговременное цитирование изменчивых, мигрирующих, новых, теряющих актуальность и даже не хранимых данных. [17].

Большинство видов повторно используемых данных, требующих больших затрат на их производство, являются специализированными базами данных. Принципы FAIR требуют, публикации достаточно подробных метаданных, определяющих такие данные и позволяющих обнаруживать их содержимое. Схемы метаданных должны быть доступными вместе с любыми ограничениями доступа к соответствующим данным. Наряду с такими видами тщательно сопровождаемых баз данных, важные данные образуются в процессе проведения различных экспериментов в виде быстро растущего разнообразия репозиториев разного назначения и форм, унификация и приведение к общему виду типов данных в которых практически трудно реализуемы. Примеры репозиториев такого вида содержатся в [15]. Это означает, что создаваемая инфраструктура должна быть более разнообразной ввиду усложнения повторного использования данных для людей и машин.

Собственно, руководящие принципы представления FAIR данных опубликованы в [15] следующим образом (учитывая важность однозначной трактовки этих определений, предпочтение отдано их публикации на языке оригинала):

- *To be Findable*: F1. (meta)data are assigned a globally unique and persistent identifier; F2. data are described with rich metadata (defined by R1 below); F3. metadata clearly and explicitly include the identifier of the data it describes; F4. (meta)data are registered or indexed in a searchable resource.
- *To be Accessible:* A1. (meta)data are retrievable by their identifier using a standardized communications protocol; A1.1 the protocol is open, free, and universally implementable; A1.2 the protocol allows for an authentication and authorization procedure, where necessary; A2. metadata are accessible, even when the data are no longer available.
- *To be Interoperable:* I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation; I2. (meta)data use vocabularies that follow FAIR principles; I3. (meta)data include qualified references to other (meta)data.
- *To be Reusable:* R1. meta(data) are richly described with a plurality of accurate and relevant attributes; R1.1. (meta)data are released with a clear and accessible data usage license; R1.2. (meta)data are associated with detailed provenance; R1.3. (meta)data meet domain-relevant community standards.

Сочетание «(мета)данные» означает, что действие соответствующего принципа распространяется как на данные, так и на метаданные. Какие-либо соображения относительно техники реализации представлений FAIR данных и метаданных были намеренно опущены в [15]. Примеры крупных биомедицинских коллекций, приведенных в этой статье, позволяют употребить интуицию для рассуждений о метаданных, которые следует использовать для поддержки различных принципов (F, A, I, R).

Вместе с тем, в самом начале опубликование принципов FAIR организацией FORCE11 в 2014 г. [16] сопровождалось предложениями не по технической реализации таких принципов, а определением необходимых характеристик данных, метаданных, сервисов и инфраструктур для того, чтобы данные были FAIR. Прежде всего, в нем введено понятие «Объекта данных» как идентифицируемой статьи данных (data item), включающей элементы данных, метаданные и идентификатор.

Например, принцип “to be interoperable” уточняется следующим образом:

- 3. Data Objects can be Interoperable only if:
- 3.1. (Meta) data is machine-actionable (https://www.force11.org/node/6062/#Annex6-9)
- 3.2. (Meta) data formats utilize shared vocabularies and/or ontologies (https://www.force11.org/node/6062/#Annex6-9)
- 3.3 (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible (https://www.force11.org/node/6062/#Annex10-11)

Здесь пункты 3.1, 3.2 и 3.3 сопровождаются ссылками на соответствующие разделы документа [16], содержащие определения, позволяющие неформально проверять, являются ли данные в некоторой конкретной инфраструктуре FAIR. Если при такой проверке ответ отрицательный, то можно дать объяснение, почему.

Учитывая, что для обеспечения охвата конкретной инфраструктуры данных средствами EOSC принципы FAIR должны поддерживаться, в ближайшее время можно ожидать появления большого числа разнообразных технических реализаций FAIR данных. В настоящей публикации даны несколько примеров начала такого движения – в том числе, в следующем параграфе рассматриваются тьюториалы, представленные на конференции в декабре 2016 г. в Амстердаме, а в разделе 5.2.5 даны краткие сведения о пилотном проекте EUDAT по исследованию возможности поддержки принципов FAIR в инфраструктуре EUDAT.

Пять тьюториалов, представленных на 9-й Международной конференции «Semantic Web Applications and Tools for Life Sciences» в Амстердаме 5 – 8 декабря 2016 г. в некоторой степени монстрировали, как FAIR принципы могут быть достигнуты с помощью средств Семантического Веба. Так, например, показано, как 18 словарей, составленных на RDF, удалось использовать для представления общих элементов метаданных и множеств их значений. В спецификациях был выделен 61 обобщенный элемент метаданных, относящихся к описанию данных для лицензирования, атрибутирования, образования версий, происхождения данных, аннотирования контента. Показано, как существующие словари можно использовать для удовлетворения требованиям представления FAIR данных при реализации индексирования, обнаружения данных, обмена данными, а также запросов и поиска при помощи SPARQL в наборах данных. Для надлежащего представления клинических данных, их интеграции и поддержки федераций данных, поддерживаемых в ряде больниц, использовались RDF, онтологии на RDF(S) и OWL, а также SPARQL. Показано, как переходить от метаданных в DICOM к их представлению в RDF. В одном из тьюториалов было показано, как от спецификации на FHIR стандарта HL7 переходить к представлению метаданных и данных в RDF.

Этот параграф завершает краткая аннотация работы, опубликованной в октябре 2016 г. и посвященной предложению по интерпретации интероперабельности в контексте FAIRness при помощи комбинации технологий Семантического Веба [18].

Авторы утверждают, что результат хорошо согласуется с принципами FAIR и рекомендуют ее в качестве эталонной реализации FAIR для случаев неинтероперабельных форматов данных в любом репозитории.

Общие рекомендации по использованию технологий Семантического Веба, собранные Н.А.Скворцоым из ряда публикаций, относящихся к поддержке принципов FAIR, выглядят следующим образом. Linked Data Platform (LDP) используется для реализации операций чтения

метаданных по URL через HTTP для получения данных в RDF. Data Catalogue Vocabulary (DCAT) используется для описания массивов данных, элементов структур данных и для распределения данных. Simple Knowledge Organization System (SKOS) – для онтологического описания записей данных. Triple Pattern Fragments (TPF) позволяет посылать посредством URL запросы по образцу, аналогичные запросам SPARQL, включающие определение триплетов с переменными. Правила RDF Modelling Language (RML) трансформируют представление оригинальных наборов данных в язык RDF.

Пока целостной реализации принципов FAIR, которую можно было бы использовать для преобразования произвольной коллекции данных в FAIR коллекцию, не существует. Пожалуй, что и подобных попыток до сих пор не было.

5.2.2 Объявление Еврокомиссии об инициативе Европейского облака, которое должно обеспечить Европе глобальное лидерство в экономике, движимой данными

19 апреля 2016 г. Еврокомиссия представила идеи плана создания облачных сервисов и глобальной инфраструктуры поддержки науки, бизнеса и общественных услуг с целью получения максимальных выгод от революции, вызванной развитием больших данных [19].

Европа является крупнейшим производителем научных данных в мире, но недостаточная и фрагментированная инфраструктура приводит к тому, что не удается использовать полный потенциал этих «больших данных». Путем поддержки и объединения существующей исследовательской инфраструктуры, Комиссия планирует создать новое Европейское облако Открытой науки (EOSC), которое предоставит 1,7 миллионам исследователей Европы и 70 миллионам профессионалов науки и техники виртуальную среду хранения, совместного и повторного использования данных через границы в междисциплинарной среде. Это будет подкреплено Европейской инфраструктурой данных, развертыванием сетей с высокой пропускной способностью, крупномасштабными средствами хранения данных и суперкомпьютерной мощностью, необходимой для эффективного доступа и обработки больших наборов данных, хранимых в облаке. Эта инфраструктура мирового класса гарантирует участие Европы в глобальной гонке по высокопроизводительным вычислениям в соответствии с ее экономическим и научным потенциалом. Цель этой инициативы заключается в создании облака Открытой науки (EOSC), чтобы сделать науку более эффективной и продуктивной и дать возможность миллионам исследователей по всему миру совместно использовать и анализировать исследовательские данные в доверенной среде независимо от технологий, дисциплин и границ. Ключевая компонента этого влияния заключается в изменении способа выполнения научных исследований по мере быстрого введения Открытой науки.

В процессе создания EOSC требуется преодолеть следующие основные технические и организационные проблемы:

- Европейские научные и деловые сообщества, общественные круги не в состоянии использовать полный потенциал данных и его трансформационное влияние на традиционные способы проведения исследований; данные, получаемые от общественно фондируемых исследований, не всегда являются открытыми (например, академическому сообществу не всегда ясны преимущества совместного использования данных);
- фрагментация инфраструктур данных, порождаемая научными и экономическими причинами, национальными границами государств, традиционными моделями управления, препятствует развитию науки, движимой данными, мультидисциплинарному международному научному сотрудничеству;

- растущие потребности в Европе в создании мирового класса высокопроизводительных компьютерных инфраструктур, ориентированных на обработку данных в науке и технике (например, для моделирования крупных технических изделий в процессе их создания, природных явлений и систем, и пр.);
- производители и потребители научных данных должны иметь возможность повторного использования данных и передовых аналитических методов, включая извлечение знаний из текстов и данных, в среде, которая должна быть не менее надежной, чем собственные средства.

В процессе создания EOSC предлагается достичь следующего:

- Европейское облако открытой науки начнет свое функционирование с федерализации существующих научных инфраструктур данных, рассредоточенных сегодня по дисциплинарному принципу и географически среди государств – членов ЕвроСоюза; в том числе будет обеспечен открытый доступ к публикациям и данным, полученным при поддержке Horizon 2020 (проекты должны производить данные, удовлетворяющие принципам FAIR) при гарантии приватности и защиты данных; изменена структура стимулирования академических, промышленных и общественных служб в реализации совместного использования данных, а также в усилении подготовки персонала в управлении данными, в их надзоре (stewardship);
- создать спецификации поддержки интероперабельности и совместного использования данных между различными дисциплинами и инфраструктурами, имея ввиду их последующую стандартизацию;
- создать настраиваемую по месту панъевропейскую структуру управления федерализацией научных инфраструктур данных и препятствованием фрагментации;
- создать облачные сервисы для поддержки Открытой науки, поддерживаемые Европейской инфраструктурой данных;
- расширить базу научных пользователей EOSC исследователями и создателями инноваций из всех областей исследований и стран членов Евросоюза, равно как из стран-партнеров и глобальных инициатив, также база пользователей EOSC будет расширена за счет общественного сектора и промышленности.

Европейская Инфраструктура Данных (EDI), реализуемая для поддержки EOSC, должна включать высокопроизводительные компьютеры, высокоскоростные сетевые средства и передовые сервисы данных (предполагается, что эти сервисы будут основаны на развитии существующих сервисов в е-инфраструктурах OpenAIRE, EUDAT, EGI, IndigoDataCloud, HelixNebula, PRACE, GÉANT) для поддержки ученых и других продвинутых пользователей от промышленности и общественного сектора. К 2022 в Евросоюзе ожидается создание суперкомпьютера эксамасштаба на основе европейской технологии, входящего в число первых трех суперкомпьютеров в мире. EDI будет работать в комбинации с национальными и региональными научными и общественными центрами данных при обеспечении должного уровня их интероперабельности.

Общественные и приватные вложения, необходимые для реализации Европейской Облачной Инициативы, оцениваются в 6.7 миллиардов евро. По оценкам Комиссии, 2 миллиарда евро будут выделены в рамках программы HORIZON 2020. Дополнительные общественные и приватные вложения составляют 4.7 миллиарда евро в течение 5 лет.

5.2.3 Проект создания Интернета FAIR данных и сервисов

«GO-FAIR» - это координированная Нидерландами международная инициатива, целью которой является реализация *Интернета FAIR данных и сервисов*, в котором аналитические инструменты будут соединены с релевантными данными, так что и инструменты и данные должны соответствовать требованиям FAIR (с деталями GO-CHAIR можно познакомиться по заявке на реализацию проекта, представленной в июне 2016 г. в агенство National Icons Нидерландов [20].

Быстро развивающаяся тенденция заключается в использовании сложной аналитики, совмещающей множество источников данных и областей знаний. Глобальные инициативы, будь то относящиеся к прецизионной и персонализированной медицине (например, Health-RI в Нидерландах), прецизионному сельскому хозяйству, логистике и демографии или охране окружающей среды, все они расширяют свои возможности за счет глобальной доступности FAIR-данных и сервисов.

Проблемы управления данными и аналитики являются очень важными в науках о жизни. Именно поэтому GO-FAIR начнет свою работу именно в этом направлении, но другие научные дисциплины также получат выгоду от открытой реализации как междисциплинарно, так и в рамках управляемого международного сотрудничества при помощи GO-FAIR. Обеспечивая ключевые компоненты Интернета FAIR данных и сервисов через систему Лабораторий открытой реализации (Open Implementation Labs) GO-FAIR, этот проект затронет все секторы глобального общества, а также послужит росту экономического потенциала.

Согласно проекту, Нидерланды позиционируют себя как инновационный центр (hub) для инициатив EOSC в Европе и инициативы COMMONS в США, которую они считают родственной.

5.2.4 Европейское облако Открытой науки

Этот раздел обзора написан на основе первого отчета Экспертной группы, учрежденной Европейской Комиссией для выработки рекомендаций по реализации Европейского облака Открытой науки (EOSC) [21]. Концепция EOSC трактуется следующим образом: это федеративная, глобально доступная среда, в которой исследователи, провайдеры инноваций, компании и граждане могут публиковать, находить и повторно использовать данные и инструменты для исследований, инноваций и образования. Такая среда должна выращиваться в Европе и за ее пределами для обеспечения того, чтобы Европейские исследования и инновации могли полностью служить созданию знаний, а также противостоять глобальным вызовам и служить экономическому процветанию Европы. Тем самым EOSC будет служить средой для Открытой науки, устраняя технические, юридические и гуманитарные барьеры в повторном использовании исследовательских данных, инструментов, а также обеспечивая доступ к сервисам, системам и потоку данных сквозь дисциплинарные, социальные и географические границы. Термин «облако» является метафорой, подчеркивающей бесшовность и общность используемых данных, программ, стандартов, экспертизы и политики, относящихся к науке, движимой данными, и к инновациям.

Среди социальных и технических вызовов на пути создания EOSC в отчете упоминаются:

- технические вызовы заключаются не столько в объеме данных, а скорее в сложности данных и аналитических процедур в различных областях;
- существование пропасти между провайдерами е-инфраструктур и специалистами в научных областях;
- фрагментация таких областей, приводящая к повторяющимся решениям и изоляции при создании и развитии исследовательских инфраструктур;

- постоянно растущие распределенные коллекции данных становятся все более немобильными, а централизация суперкомпьютеров самих по себе становится недостаточной для поддержки в сущности федеративного и распределенного мета-анализа и обучения;
- необходимые для создания первого поколения EOSC компоненты существуют, однако они разбросаны и теряются среди 28 стран – членов Евросоюза, а также среди разнообразных сообществ.

К побудительным факторам создания EOSC как части Открытой науки относятся следующие: потребность в новых способах коммуникаций в научном мире (с упором на действенность машин); осознание и признание необходимости совместного повторного использования данных; потребность в инновационных схемах устойчивой поддержки исследовательских инфраструктур; требования мультидисциплинарного сотрудничества; EOSC должно играть роль эко-системы над инфраструктурами; индикаторы эффективности должны быть определены.

Облако EOSC должно быть интероперабельным с Интернетом FAIR данных и сервисов и быть доступной инфраструктурой для проведения исследований и введения инноваций. В нее входят также опыт людей, ресурсы, стандарты, практические решения и нижележащие инфраструктуры. Наряду с управлением данными, в EOSC должен поддерживаться надзор (stewardship) данных, обеспечивающий должное качество данных и препятствующий их утрате.

Толкование открытости уточняется так, что ‘open should not be confused with free’ (хотя ученые могут ожидать, что данные и связанные с ними сервисы могут использоваться ими бесплатно в противоположность коммерческому подходу). Открытость скорее относится к доступности всех элементов EOSC при надлежащих и хорошо определенных условиях.

Ряд тенденций, характерных для Открытой науки, важно учитывать при создании EOSC. Так, новые способы научных коммуникаций заключаются в вовлечении в этот процесс главных ассистентов исследований – машин, генерирующих и обрабатывающих данные. В Европе, где в академической среде до сих пор важность экспертов по данным была сильно недооцененной, отсутствие экспертизы данных рассматривается среди основных рисков утраты ведущих позиций в науке.

Развитие кросс-дисциплинарного сотрудничества является важной тенденцией ввиду того, что участились случаи использования сырых данных и аналитических средств из других, весьма отдаленных дисциплин. В то же время, учитывая отсутствие стандартов приемлемых метаданных и специальных поисковых инструментов, трудно обвинять исследователей в изобретении колеса. Использование текстов и техники data mining будет важным в EOSC для кросс-дисциплинарного использования.

Важной тенденцией является развитие сложных экосистем инфраструктур. Рост объемов данных приводит к необходимости перемещать потоки работ (на основе техники процессных виртуальных машин) к данным, а не данных к потокам работ.

Развитие понимания в машинах, рассматриваемых в качестве ассистентов. Выделяются два вида существенно различных функций таких ассистентов (распознавание образов – исключительно машинная функция; функция, связанная с семантикой и отображением терминов и идентификаторов, отслеживанием происхождения данных, оценкой результатов и их интерпретации, требует привлечения когнитивных подходов). Обеспечение действенности машин в этой второй функции потребует серьезных усилий.

Примеры рекомендаций отчета для действий Комиссии на подготовительном этапе, дающие представление о проблемах и масштабах инициативы EOSC (хотя ряд рекомендаций напоминают лозунги):

- Позиционировать EOSC как вклад Евросоюза, поддержанный открытыми протоколами, в Интернет FAIR данных и сервисов.
- Действовать в предположении, что все данные в EOSC (в действительности, все Research Objects) должны быть FAIR.
- В области действия принципов FAIR стандарты и протоколы должны быть ограничены абсолютно минимальными решениями для уменьшения риска того, что их будущее развитие потребует адаптации протоколов.
- Провести быстрое прототипирование и создать эталонные реализации критических элементов подготовительного этапа EOSC при поддержке Horizon 2020 с тем, чтобы уже в 2017 году образцы рабочих сред могли быть реализованы в ключевых дисциплинах в ведущих государствах - членах Евросоюза с целью их последующего распространения в других сообществах и странах.
- Существующие и будущие инструменты планирования и финансирования исследований, включая Horizon 2020, должны поддерживать только те проекты, которые надлежащим образом относятся к вопросам организации надзора (Stewardship) для открытых данных. Проекты создания изолированных инфраструктур данных, не требующие соблюдения принципов FAIR данных, не предусматривающие вклада в видение EOSC как всеобщей инфраструктуры для оперирования данными, не должны считаться приемлемыми для финансовой поддержки.

5.2.5 Подход EUDAT к интероперабельности FAIR данных

Для поддержки публикации данных согласно принципов FAIR и обеспечения их обнаружения, доступа, интероперабельности и повторного использования, разрабатывается набор инструментов (FAIR) данных, в том числе FAIR Data Point (FDP), представляющий собой программный слой над наборами даных, представляющий их в виде взаимосвязываемых (inter-linkable) FAIR данных [22]. FDP предоставляет информацию о доступных наборах данных с точки зрения их метаданных, а также обеспечивает доступ к самим данным в интероперабельном формате. В рамках этого пилотного проекта будет выяснено, возможно ли расширить функциональность существующих сервисов EUDAT функциями FDP или же необходимо разработать новый сервис, основанный на FDP.

Стоит подчеркнуть, что подобный сервис будет разработан впервые, поскольку на сегодняшний день для широкой научной общественности нет ни одного сервиса Semantic Web для поддержки репозиториев. Более того, FDP проектируется так, чтобы поддерживать цитирование данных и вести статистику доступа к данным, что позволит оценивать эффективность каждого развернутого FDP. Целью данного пилотного проекта является реализация и развертывание FDP, используя комбинацию существующих стандартов и каркасов Семантического Веба для разработки клиентской части (front-end) и (существующие или новые) сервисы EUDAT – для разработки серверной части (back-end). FDP обеспечивает доступ к данным и метаданных, используя REST-API, соответствующих спецификации W3C Linked Data Platform (Платформа связанных данных). Функциональность, которую предоставляет FDP, расширяет возможности обнаружения, доступа, интероперабельности и многократного использования семантически богатых исследовательских данных. В первую очередь, проект нацелен на такие сервисы EUDAT, как B2Safe и B2Share. Его

конечная цель – достичь соответствия этих сервисов требованиям FAIR Data Point, придерживаясь, таким образом, принципов FAIR, и предоставляя данные и метаданные в стиле FAIR.

По завершении проекта, ожидается демонстрация возможности образования и преимуществ провайдера сервисов крупномасштабных репозиториев данных подобного EUDAT, позволяющего представлять опубликованные наборы данных согласно принципов FAIR. Разработка данной инфраструктуры [22] также является частью подготовки к предстоящим требованиям European Open Science Cloud.

6. Программа реализации EOSC

В конце 2017 года Еврокомиссия объявила о планах и финансировании реализации EOSC в 2018 – 2020 годах. Этому предшествовал ряд действий.

Из документов, опубликованных Еврокомиссией в 2016 г., здесь заслуживает упоминания документ (http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52016DC0178), направленный в Европарламент 19 апреля 2016 г. для сообщения основных положений позиции Еврокомиссии по отношению к Европейской Облачной Инициативе: "European Cloud Initiative - Building a competitive data and knowledge economy in Europe", которые, по замыслу авторов, должны обеспечить Европе достойное место в глобальной экономике, движимой данными.

26 января 2017 г. опубликован отчет о пленарном заседании Комитета Европарламента по промышленности, исследованиям и энергии для обсуждения Предложения Комиссии по Европейской Облачной Инициативе (European Cloud Initiative), требующего одобрения Парламентом. Более полное название инициативы звучит так: 'European Cloud Initiative – Building a competitive data and knowledge economy in Europe'. Таким образом, в комитете Европарламента речь шла о потенциале открытой науки и облачных вычислений как части цифровой экономики Европы (http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&reference=A8-2017-0006&language=EN). Это формальный документ, структура которого соответствует требованиям Европарламента: вначале следуют ссылки на публикации и предшествующие резолюции их рассмотрения в различных организациях Евросоюза, далее следуют уточнения терминов и предлагаемых решений, и, наконец, следуют формулировки предложений, требующих утверждения парламентом. Структура последнего раздела следующая: Общие предложения; Облако открытой науки; Совместное использование открытых и исследовательских данных; Извлечение информации из текстов и данных; Защита данных, фундаментальные права и безопасность данных. Общее число предложений в этом разделе 108. Далее в отчете следуют замечания и рекомендации комитетов, сформулированные ранее, и, наконец, результат голосования каждого из профильных комитетов (распределение голосов в ответственном за предмет комитете: ЗА – 46, ПРОТИВ – 7).

Примеры предложений из раздела "Облако открытой науки":

39. Stresses that the Open Science Cloud Initiative should lead to a trusted cloud for all: scientists, businesses and public services;

40. Notes that there is a necessity to foster an open, trusted collaborative platform for the management, analysis, sharing, reuse and preservation of research data on which innovative services can be developed and delivered under certain terms and conditions;

43. Asks the Commission to ensure that all scientific research and data produced by the Horizon 2020 programme is open by default, and asks the Member States to adapt their

national research programmes accordingly;

48. Welcomes the fact that the Cloud Initiative focuses on building high-bandwidth networks, large-scale storage facilities, high-performance computing and a European big data ecosystem;

57. Stresses that the use of open standards, and free and open-source software, are especially important in guaranteeing the necessary transparency about how personal and other sensitive types of data are in fact being protected;

67. Calls on the Commission to take the lead in promoting intersectoral, cross-lingual and cross-border interoperability and cloud standards, and in supporting privacy-friendly, reliable, secure and energy-efficient cloud services as an integral part of a common strategy focusing on maximising the opportunities to develop standards that have the capacity of becoming worldwide standards;

76. Supports the Commission's intention to remove barriers, especially technical and legal ones, to the free movement of data and data services, to remove as well disproportionate data localisation requirements, and to promote the interoperability of data by linking the European Cloud Initiative to the Free Flow of Data Initiative.

26 октября 2017 г. Еврокомиссией опубликована декларация EOSC (http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud), начало формирования которой было положено на саммите EOSC 12 июня 2017 г. В этот же день опубликован EOSC declaration action list, в котором, например, упоминается FAIR Data Action Plan 2018 – 2020. Декларация содержит принципы реализации EOSC (http://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none), одобренные представителями организаций, поддерживающих декларацию (список около 100 организаций-участников от 24 ноября 2017 г. опубликован здесь: http://ec.europa.eu/research/openscience/pdf/list_of_institutions_endorsing_the_eosc_declaration.pdf#view=fit&pagemode=none).

Декларация подчеркивает, что реализация EOSC это не проект, а процесс, по природе итеративный и основанный на постоянном обучении и взаимной подстройке. Декларация включает следующие разделы:

- Культура данных и данные, обладающие свойствами FAIR;
- Сервисы и архитектура организации исследовательских данных;
- Управление и финансирование.

Наконец, 27 октября 2017 г. опубликовано сообщение Еврокомиссии (http://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-infrastructures_en.pdf) о конкурсе грантов на проекты программы H2020 по реализации Европейских исследовательских инфраструктур в период 2018 – 2020 г., включающем проекты реализации EOSC в этот же период. Начало конкурсов грантов на шесть различных проектов реализации EOSC разнесены по времени с 5 декабря 2017 г. до 14 ноября 2018 г. Deadlines для подачи предложений соответственно разбросаны по времени с 22 марта 2018 г. до 20 марта 2019 г. Бюджеты этих проектов определены. Суммарный объем финансирования составляет 142 миллиона евро в 2018 г., 45.5 миллионов евро в 2019 г. и 84 миллиона евро в 2020 г. Максимальное время для подписания соглашения по гранту – 8 месяцев от deadline для подачи предложения.

Меньше всего информации имеется по третьему проекту из шести. Известна только тема проекта и то, что он начнется в 2020 году:

INFRAEOSC-03-2020: Integration and consolidation of pan-European access mechanisms to public e-infrastructures and commercial services through the EOSC hub

Ввиду того, что этот проект ориентирован на применение работающего Европейского облака, такое обособленное отношение к нему представляется оправданным.

Далее следует краткая аннотация проектов, по которым имеются определения их содержания в сообщении Еврокомиссии от 27 октября.

INFRAEOSC-01-2018: Access to commercial services through the EOSC hub

Два вида сервисов должны быть рассмотрены: а) товарный тип коммерческих цифровых сервисов необходимых в междисциплинарных исследовательских делах, включающий, например, облачные сервисы (хранения, вычислений, приложений), софтверные лицензии, симуляционный инструментарий, тулы виртуализации и поддержки совместной деятельности; б) различные виды коммерческих сервисов данных (космической и земной природы); например, коммерческие сервисы наблюдения Земли из космоса, включающие прежде всего предоставление информации, основанной на открытых данных проекта Коперник, доставляемых при помощи платформ 'DIAS' и ее умной интеграции в каталог сервисов EOSC.

При этом необходимо опираться на анализ, определяющий как различные нужды исследователей могут быть агрегированы с дополнительными коммерческими сервисами для максимизации вклада от инструментария уже представленного каталогом сервисов EOSC для поддержки междисциплинарных исследований. Кроме того, нужно обеспечить техническую интеграцию коммерческих сервисов в каталог сервисов EOSC, включая управление правами доступа и установление приемлемых правовых соглашений, а также определение надлежащих соображений по использованию сервисов и их интероперабельности.

Следует также понимать, что новая панъевропейская модель исследовательских данных и сервисов должна быть создана на основе EOSC хаба (панъевропейского механизма доступа к сервисам, предоставляемого на национальном, региональном или институциональном уровне). Любопытно, что в сноске на стр. 15 сказано, что механизм EOSC хаба будет основан на предложении, ожидающемся от проекта, который победит на конкурсе EINFRA-12-2017.

INFRAEOSC-02-2019: Prototyping new innovative services

На основе возможностей хаба EOSC, инновационные сервисы должны быть созданы с тем, чтобы учитывать релевантные аспекты цикла исследовательских данных (от начального состояния до публикации, курирования, сохранения и повторного использования). Подчеркивается эволюционный характер создания инновационных сервисов (сначала они могут удовлетворять потребностям конкретного научного сообщества, но к концу проекта они должны удовлетворять требованиям междисциплинарного исследования). Аналогичная эволюция допускается в отношении TRL, который должен быть к концу проекта не ниже 8 для систем и технологий, используемых предлагаемыми сервисами. Также рекомендуется использовать предложения проектов-победителей конкурса EINFRA-12-2017, относящиеся к хабу EOSC.

INFRAEOSC-04-2018: Connecting ESFRI infrastructures through Cluster projects

В рамках этой тематики следует использовать кластеризацию проектов и вех ESFRI в следующих областях исследований: Биомедицинские науки; Окружающая среда и науки о Земле; Физика и ее средства аналитики; Социальные и гуманитарные направления; Астрономия; Энергетика. Каждая инфраструктура должна участвовать только в одном кластере. Предложения должны включать подходы к обслуживанию данных во включаемых инфраструктурах следуя принципам FAIR согласованно с целями Открытой Науки. При этом должны рассматриваться определения специфических для каждой области политик данных, вопросов интероперабельности и права, которые влияют на трансграничное оперирование данными в рамках географических и тематических границ. Предложения должны рассматривать вопросы синергетики и дополнительности при оперировании данными между различными инфраструктурами, включая проблемы интеграции и интероперабельности данных в EOSC.

Предлагаемые консорциумы должны включать ключевых участников вовлеченных инфраструктур и их правовые подразделения. Ожидаемое влияние проекта должно включать область целей Открытой Науки, охватывая такие вопросы как междисциплинарную открытую инновационную среду для поддержки данных, знаний, сервисов, образование глобальных стандартов, онтологий и интероперабельности научных данных, принятие общих подходов к жизненному циклу управления данными, и др.

INFRAEOSC-05-2018-2019: Support to the EOSC Governance

Главной целью этих проектов является создание операционной структуры для поддержки всеобъемлющего управления EOSC, включая координацию релевантных национальных инициатив. Операционная структура представлена тремя направлениями (каждое из которых будет поддержано отдельным подпроектом):

1) Обеспечение координационной структуры для поддержки деятельности Исполнительного Совета EOSC, отвечающего за реализацию EOSC. В этом направлении выделяются деятельности по Координации и Поддержке. Перечислены в общем виде предполагаемые виды действий (включая поддержку процессов принятия решений при реализации ключевых функций EOSC; поддержку создания инновационных моделей внедрения таких функций; набор методов и правил для использования сообществами пользователей, провайдеров инфраструктур для совместного проектирования и внедрения облачных решений и сервисов, доступных конечным пользователям при помощи хаба EOSC; обеспечение заключения соглашений со странами, не входящими в ЕвроСоюз, относительно политики и технологических разработок, совместимых с EOSC).

2) Координацию различных национальных инициатив/инфраструктур данных/e-инфраструктур и их федерализации в составе EOSC. В этом направлении (в него включаются Исследовательские и Инновационные Деятельности), наряду с его детализацией, рекомендуется при развитии стандартов сервисов и интероперабельности на различных уровнях использовать документ “European Interoperability Framework - Implementation Strategy”, принятый в марте 2017 г. Этот документ связан с другим большим изданием Еврокомиссии, озаглавленным “New European Interoperability Framework” (ISBN 978-92-79-63756-8; doi:10.2799/78681).

3) Стимулирование культуры обращения с данными FAIR и внедрение достойных практик, придающих данным свойства FAIR. В предложениях рекомендуется опираться на координацию глобальных инициатив (указаны GO-FAIR, а также сообщества CODATA, RDA, WDS), а также инициатив в странах ЕвроСоюза или ассоциированных с ним. Кроме того, рекомендуется развивать

Европейский каркас компетенции в сфере data science путем его представления в академических куррикулумах (подобно проекту EDISON).

INFRAEOSC-06-2019-2020: Enhancing the EOSC portal and connecting thematic clouds

Основная задача проекта – поддержка портала EOSC. Эта тема ориентирована на обеспечение полнофункционального, устойчивого и полного пользовательского интерфейса, который мог бы служить в качестве универсальной точки входа для сервисов EOSC. Предложения рекомендуется базировать на результатах проекта eInfraCentral и обеспечить дополнительную поддержку реализации хаба EOSC для окончательного формирования и оптимизации функций и интерфейсов портала EOSC. Ожидаемые проблемы связаны с необходимостью обеспечения открытой среды для пользователей, ведущих исследования в произвольных научных областях вдоль всего жизненного цикла научных данных.

Дополнительная информация, имеющая отношение к программе реализации EOSC, приведена далее. Заметным является повышение плотности событий.

Список исследовательских инфраструктур, связанных с Horizon 2020 и предоставляющих свободный доступ к данным, опубликован в феврале 2017 г. в документе «Research Infrastructures offering free Access with EU support». Список включат проекты в следующих областях: Биомедицинские науки; Окружающая среда и науки о Земле; Физические науки; Социальные и гуманитарные направления; Энергетика; Материальные науки и средства аналитики; Математика и ICT.

В мае 2017 г. в 27 выпуске новостного издания Inspired, издаваемого EGI сообществом (www.egi.eu), наряду с представлением сервисов EGI для открытой науки, анонсирован двухлетний пилотный проект (http://eoscpilot.eu/), предназначенный для формирования оснований EOSC, в котором участвуют 33 Европейских партнера. В июне 2017 г. EGI консорциум опубликовал брошюру EGI Use Cases. В ней наряду с описанием интересных примеров применений инфраструктур данных в науке кратко описано EGI Federated Cloud, которое объединяет приватные академические облака и виртуализованные ресурсы на основе открытых стандартов. Основным является абстрактный Cloud Management Framework, поддерживающий множество облачных интерфейсов.

В ноябре 2017 г. EGI консорциум опубликовал брошюру «EGI support for Research Infrastructures» (https://www.egi.eu/wp-content/uploads/2017/11/EGI-RIs.pdf), содержащей аннотации избранных проектов из 31 исследовательских инфраструктур, поддерживаемых EGI, примеров совместных проектов в областях Науки об окружающей среде (5 проектов); Гуманитарные исследования (1 проект); Физика и астрономия (2 проекта); Биомедицинские науки (4 проекта). EGI представляет собой федерацию из 300 центров данных и вычислений (полный список центров данных содержится здесь: egi.eu/federation/data-centres/) и 21 облачных провайдеров. Полный список EGI сервисов можно найти здесь: go.egi.eu/ext, go.egi.eu/int и https://www.egi.eu/services.

В то же время EGI выступил с одобрением декларации EOSC (https://www.egi.eu/news/egi-endorses-the-eosc-declaration/). Информация о поддерживаемых EGI инфраструктурах и некоторых проблемах реализации EOSC (включая создание FAIR сервисов и хаба EJSC) содержится в материалах конференции DI4R (https://www.egi.eu/news/egi-highlights-at-di4r-2017/).

В январе 2018 г. EUDAT организует конференцию “Putting the EOSC vision into practice”, программа которой содержится по адресу: (https://eudat.eu/eudat-conference-2018-programme).

7. Проект «Национальный сервис данных» (NDS) в США

Национальный сервис данных (National Data Service, NDS), разрабатываемый в настоящее время в США, должен обеспечить стандартный набор услуг для хранения, совместного использования, публикации, размещения и верификации данных во всех дисциплинах и сообществах, однако, он должен быть построен над существующими инфраструктурами, уже используемыми соответствующими сообществами. NDS функционирует как консорциум (в соответствии с уставом [23]), разрабатывающий открытую среду распределенных, интероперабельных и интегрированных сервисов национального масштаба. В соответствии с этим подходом, NDS берет на себя ответственность адаптировать существующие или разработать необходимые новые сервисы, сосредоточившись на следующих пяти ключевых услугах:

- обанаружение данных, созданных или хранимых учеными или исследователями;хранение постоянных копий курируемых данных и связанных с ними метаданных для архивирования, совместного использования, публикации и других целей;
- доступ к данным при помощи репозиториев и других мест их хранения;
- связывание данных с другими данными, их публикациями, а также возможность повторного использования;
- обработка и анализ данных для развития науки, линий поведения и инноваций.

Устав NDS [23] является руководством по управлению операциями NDS и Консорциума Национального сервиса данных (National Data Service Consortitum, NDSC). Если Национальный сервис данных – это организация, предоставляющая совокупность услуг, то Консорциум Национального сервиса данных – это более широкое сообщество заинтересованных лиц, включающее физических и юридических лиц.

По сравнению с EOSC, NDS представляется более индустриальным проектом: здесь (может быть, пока) нет стремления к введению семантических метаданных для поддержки принципов FAIR. Поэтому важным компонентом NDS является NDS Labs, который рассматривается, как "игровая площадка" для оценки, разработки и интеграции технологии управления данными. Это среда, где разработчики могут проектировать инструменты и тестировать новые возможности, используемые при создании каркаса и сервисов NDS. Среда предоставляет быстрый доступ к инкапсулированным средствам и сервисам управления данными таким образом, что они могут быть быстро развернуты с целью оценки или разработки. Она позволяет разработчику или небольшой команде разработчиков проверять новые идеи, проектировать новые сервисы или объединять существующие приложения как часть создания экосистемы NDS. Таким образом, просто проводить следующие операции:

- Выполнять хостинг сервисов
- Получать доступ к эластичным вычислительным ресурсам (виртуальным машинам)
- Получать доступ к хранилищам данных
- Находить доступные средства и сервисы управления данными
- Сравнивать и проводить оценку различных технологий
- Разворачивать тестовые экземпляры программ
- Выполнять облачную разработку ПО
- Публиковать/предоставлять для совместного использования инструментов одних разработчиков другими участниками проекта.

Кроме этого, NDS Lab предоставляет оборудование для выполнения пилотных проектов, а также обеспечивает совместную поддержку деятельности по разработке программного обеспечения. Персонал, обеспечивающий поддержку коллабораций, работает с проектами киберинфраструктур данных таких организаций, как NSF, NIH, NIST, DOE и др., и готов предоставить свой опыт и знания в области технологий обработки и анализа данных для поддержки разработки в рамках различных пилотных проектов.

Перечень технологических компонентов NSF, взятых из различных программ NSF (таких как, DIBBs, DataNetm EarthCube) представлен в таблице на веб-странице [24]. Строки этой таблицы содержат гиперссылки на соответствующие компоненты, а столбцы соответствуют функциям, которые разрабатываютя NDS. Значения в ячейках соответствуют степени использования функции: Potential – потенциально рассматривается к использованию в NDS, Currently – в настоящее время ведется реализации функции в NDS, Utilized – функция полностью реализована в NDS США.

На сайте CINERGI проекта EarthCube (https://www.earthcube.org/group/cinergi) содержатся различные ссылки, касающиеся проблем интероперабельности в области наук о Земле, с которыми часто сталкиваются при поиске и интерпретации междисциплинарных данных, полученных из различных источников. CINERGI может снизить нагрузку при поиске, интерпретации и оценке пригодности к использованию различных типов информационных ресурсов из различных областей наук о Земле. В рамках ряда проектов из области наук о Земле – по геохимии, гидрологии, океаническим исследованиям, экологии и другим областям – были созданы тематические хранилища данных и каталоги метаданных. CINERGI позволяет получить к ним доступ через единый стандартизованный интерфейс и улучшает описания метаданных, чтобы сделать процесс поиска данных более единообразным и менее времязатратным. Веб-страница NDS, относящаяся к технологиям EarthCube (https://nationaldataservice.atlassian.net/wiki/display/NDSC/EarthCube+Technology+Components), содержит следующий список возможностей, относящихся к метаданным, которые, вероятно, поддерживаются CINERGI:

- Каталоги и регистры
- Интероперабельные кросс-дисциплинарные онтологии
- Средства для исследования, валидации, проверки и отображения данных
- Брокерные сервисы данных (метаданных), позволяющие переходить от одного стандарта к другому
- Брокерные сервисы интерфейсов, управляющие доступом к данным в различном формате и через различные протоколы
- Применение социальных сетей на профессиональном уровне для поддержки обмена знаниями между различными группами ученых в области информационных технологий и наук о Земле.

8. Альянс создания социальных и технических инфраструктур поддержки открытого совместного доступа к исследовательским данным (RDA)

Альянс исследовательских данных (RDA) был образован для поддержки совместного использования данных сквозь барьеры в 2013 г. Ядро организаторов включало Европейскую комиссию, National Science Foundation, NIST, Министерство инноваций Австралии. К ноябрю 2016 г. число членов альянса достигло 4500 из 115 стран. В рамках альянса образовано большое число рабочих групп и групп по интересам. Помимо разработки и принятия инфраструктурных решений,

в задачи RDA входит ускорение роста сплоченного сообщества, объединяющего спонсоров в рамках конкретных областей исследований, национальных, географических и возрастных границ. Дважды в год организуются пленарные совещания RDA в различных местах мира. Например, в марте 2015 г. на совещании в Сан-Диего рассматривались крупномасштабные инфраструктурные проекты организации и анализа данных (включая EUDAT, DataOne, CLARIN, Supercomputing and Big Data, ELIXIR, NDS и др.).

В качестве примера результатов RDA, связанных с настоящим обзором, дана ссылка на работу, которую RDA ведет по стандартам и моделям метаданных в различных ОИИД [25]. Целью этой работы является создание открытого справочника стандартов метаданных, применимых к научным данным, для использования при создании и анализе разнообразных инфраструктур. В редакции этого справочника на конец 2016 г. включены следующие разделы: Arts and Humanities, Engineering, Life Sciences, Physical Sciences and Mathematics, Social and Behavioral Sciences, General Research Data. Пример одного из стандартов: Core Scientific Metadata Model (CSMD) [26]. Он обобщает ряд научных дисциплин, особенно тех, которые можно отнести к «структурированным наукам» (таким как химия, материаловедение, науки о Земле, биохимия). Модель, лежащая в основе этого стандарта, организована вокруг понятия «исследование», в рамках которого проводится ряд экспериментов, наблюдений, анализов, симуляций, и пр. CSMD ориентирован на поддержку данных в рамках потоков работ. Модель ориентирована на иерархию структуры научного поиска: программы, проекты, исследования. Вот гиперссылки на документы, сопровождающие CSMD: A short introduction to CSMD , CSMD 4.0 Reference Document (HTML), CSMD OWL Ontology in RDF Turtle format, CSMD 4.0 Reference Document (PDF).

9. Примеры методов и средств концептуализации конкретных предметных областей

Астрономия. В рамках международного альянса виртуальной обсерватории (IVOA) развиваются подходы концептуализации предметной области астрономии и использования её в спецификациях, связываемых с данными. Большинством астрономического сообщества используются подходы к аннотированию атрибутов баз данных и каталогов неформальными семантическими спецификациями. До сих пор наиболее распространена специализированная система классификации на основе универсальных концептуальных дескрипторов (UCD), не гарантирующая однозначной интерпретации спецификаций машиной и человеком. Модели общеупотребимых областей, таких как фотометрия (PhotDM) и астрометрия (STC), называемые «моделями данных», также используются не в качестве схем для представления данных, а для аннотирования данных в их оригинальной структуре с помощью указания пути до элемента в схеме, соответствующего семантике атрибута. Смысл каждого элемента схем определён вербально. Существует также несколько известных онтологий, созданных на основе словарей, однако используются они на практике редко. В рамках настоящего проекта проведены исследования по концептуализации предметной области с использованием формальных онтологий и объектных концептуальных схем и показано преимущество повторного использования этих средств при решении различных задач [27]. Для передачи данных и метаданных широко используется формат FITS, включающий в себя возможность одновременной сериализации изображений, табличных данных и метаданных с набором параметров. Помимо этого, используются XML-форматы описания явлений (VO Event) и табличных данных (VOTable). Обеспечение астрономии данными во многом возложено на консолидацию зеркал обзоров, каталогов и баз данных в сервисах под унифицированными интерфейсами (ADS, BDB, INES, SIMBAD, VisieR, VALD, WEPDB). Современные требования к масштабируемости обработки данных приводят к выводам о необходимости

локализованной обработки в местах расположения данных и использования инфраструктур исследовательских данных.

Материаловедение. Особенностью предметной области материаловедения является акцент на выражении физических свойств материалов в связи с их структурными особенностями. В данной области наработаны подходы к созданию хорошо структурированных концептуальных спецификаций, которые используются для представления и передачи знаний о материалах. Одним из распространённых средств спецификации свойств материалов является язык MatML, определяющий XML-схему для описания химических, механических, термальных свойств материалов. Отмечалось, что данный язык не имеет чётких определений элементов. Однако для совместного использования с MatML созданы онтологические спецификации Materials Ontology, которые связывают с языком определённую семантику. Онтология определена на языке OWL, для описания метаданных используется язык RDF [28].

Науки о Земле. В науках о Земле объединяется множество связанных друг с другом ОИИД. В исследовании распространены как подходы анализа данных наблюдений, так и моделирования явлений. Необходимость в концептуальном осмыслении такой широкой области знаний ставится во главу угла давно. Однако средства концептуального описания, в основном, заимствовались из словарей и не имели достаточно формального описания (например, онтология SWEET). Инициатива GeoSemantics в рамках инфраструктуры данных EarthCube направлена на преодоление семантической неоднородности наработанных моделей и коллекций данных. Она представляет каркас для аннотирования и интеграции информационных ресурсов, основанный на технологиях открытых данных, и реализует сервисы работы с интегрированными неоднородными контролируемыми словарями (KIS), семантического аннотирования ресурсов (SAS) метаданными пространственно-временного контекста и происхождения, семантического поиска ресурсов (KDS) и потока работ (RAS) для согласования атрибутов разных информационных ресурсов [29].

10. Определение use cases для решения задач, требующих совместного использования данных из инфраструктур Европы и России, в конкретных ОИИД

Мотивация подготовки use cases и требования к их описанию заключаются в следующем:

- Необходимы аннотации use cases, ориентированные не столько на специалистов в конкретных ОИИД, а на использование архитекторами систем при создании ими новых инфраструктур поддержки open science и совместного доступа к данным. Образец use case содержит сценарий решения задачи, показывающий шаги решения, данные и сервисы, используемые на каждом шаге. К этому следует добавить ссылки на конкретные коллекции данных или на проекты проектируемых инструментов их получения, а также краткие пояснения функций сервисов.
- Важно, чтобы в каждом use case была показана целесообразность совместного доступа к данным в коллекциях России и в Европейских проектах исследовательских инфраструктур, планируемых к завершению в начале 20х годов 21 века.
- Нужно показать, что наши задачи потребуются после 2020 г. для проведения исследований с данными, которые будут производиться тогда. При этом существенно показать, что в России также будут данные, которые могут быть полезными для использования в Европейских исследованиях.
- Предлагаемые задачи должны быть ориентированы на перспективу (прежде всего заданием данных, которые будут актуальны после 2020 г.). Они должны показывать, что Европейские

программы о переходе к open science, к созданию соответствующих инфраструктур (типа EOSC), обеспечивающих совместный доступ к данным, к введению новых инструментов для получения данных очень важны для России.

Далее следуют описания use cases, подготовленные участниками проекта, в 2016 году.

Астрономия

А1. Поиск компонент астрофизических транзиентов различной природы

В общем случае задача формулируется и может быть формализована как задача поиска и исследования всех обстоятельств, сопровождающих астрофизические транзиенты. К классу таких событий относятся, по крайней мере, космические гамма-всплески (GRB), быстрые радиовсплески (FRB), одиночные нейтрино высоких энергий (IceCube sources), всплески гравитационного излучения (GW). GRB регистрируются космическими обсерваториями Swift (Gehrels et al. 2004), Fermi (Meegan et al. 2009), INTEGRAL (Winkler et al. 2003) и в будущем – SVOM (Götz et al. 2009). Точность локализации GRB составляет от нескольких угловых минут до нескольких градусов. GW детектируются наземными экспериментальными установками LIGO (https://www.ligo.caltech.edu/) и Virgo (http://www.virgo-gw.eu/). Точность локализации составляет сотни квадратных градусов. Источниками рассылки сообщений в режиме реального времени о произошедших событиях является платформа BACODINE (http://gcn.gsfc.nasa.gov/improvements.html). Исследование обстоятельств, предшествующих, сопровождающих и следующих за транзиентным событием состоит в поиске компонент события в различных диапазонах электромагнитного излучения (1) в режиме реального времени, (2) дополнительных наблюдениях и (3) в поиске в архивных данных.

1. Для всех транзиентов, производится наблюдения областей локализации в оптическом и радиодиапазонах наземными телескопами с целью поиска новых точечных транзиентных или переменных источников. В России - это сети телескопов МАСТЕР, ИСОН и ИКИ РАН. Значимость результатов работы этих сетей увеличивается со временем, результаты сообщаются в публикациях реально времени (http://gcn.gsfc.nasa.gov). Наблюдения вводимого в строй обзорного телескопа АЗТ-33ВМ (ИСЗФ, Россия, п. Монды) с апертурой 1.5м и широким полем зрения будут востребованы при поиске транзиентов.

2. Дополнительные наблюдения необходимы для целенаправленного всестороннего исследования, такие наблюдений проводятся сетью ИКИ РАН и телескопами САО РАН (https://www.sao.ru/). Как правило, это глубокие повторяющиеся наблюдения одной и той же области. Они необходимы для исследования свойств послесвечения, сверхновой, родительской галактики. Архивные данные наблюдений телескопами САО (Цейс-1000 и БТА) доступны.

3. Важной задачей поиска всех обстоятельств транзиентов является исследование архивных данных (3.1) совпадающих во времени с началом транзиентного события и (3.2) исследования места локализации точечного источника, полученных до, во время и после транзиентного события.

3.1. Производится поиск всплеска излучения в других экспериментах, в частности, в данных всенаправленных космических гамма-детекторов (SPI-ACS/INTEGRAL, Konus-Wind, БДРГ/Ломоносов). База данных Большой Сканирующей Антенны (БСА, Пущино, http://www.prao.ru/radiotelescopes/in_list_BSA.html) для радиодиапазона 110 МГц является источником поиска радио-транзиентов, сопровождающих GRB и GW, в настоящее время есть ограниченный доступ к базе данных.

3.2. Исследование места локализации точечного источника в архивных данных позволяет уменьшить объем дополнительных наблюдений. Исследование проводится в каталогах и исходных данных каталогов - изображений. Наиболее востребованный каталог для исследований свойств родительских галактик – SDSS (http://www.sdss.org/). Кросс-корреляция различных каталогов позволяет выделить калибровочные звезды, необходимые для проведения фотометрии оптических транзиентов. Несколько российских экспериментов могут быть самостоятельным источником сообщений об обнаружении транзиентов в реальном времени. БСА может быть источником сообщения о радио-транзиентах типа FRB. Реализация выделения высокоэнергичных нейтрино в режиме реального времени на установках Баксанской Нейтринной Обсерватории (БНО, http://www.inr.ru/rus/bno/lbpst.html) может быть самостоятельным источником сообщений. База данных БНО для архивного поиска нейтрино от источников GRB и GW (3.1). В сети МАСТЕР реализована отправка сообщений об оптических транзиентах различной природы. Несомненно, строящийся обзорный телескоп LSST (https://www.lsst.org/) станет основным источником оперативных сообщений об оптических транзиентах, а также неисчерпаемым источником архивных данных.

А2. Исследование астрофизических характеристик звёзд и межзвёздной среды по фотометрическим данным

Оценка таких астрофизических параметров, как температура, радиус, масса, металличность, ускорение силы тяжести для звёзд нашей Галактики требует использования данных об их спектрах в широком диапазоне излучения. Данные о спектрах большинства наблюдаемых звёзд восстанавливаются по фотометрическим данным, снятым во всех возможных диапазонах от радио до ультрафиолетового. Потоковые и обновляемые данные многополосной фотометрии, получаемые от современных инструментов SDSS, LSST (с 2021 г.), UKIDSS, Pan-STARRS (США, Европа), охватывают сведения о блеске звёзд в оптическом и ближнем инфракрасном диапазонах. Исторические данные за всё время наблюдения из каталогов GALEX, 2MASS дополняют их фотометрическими наблюдениями в инфракрасном и ультрафиолетовом диапазонах. Выборки данных перечисленных обзоров и каталогов, связанные с определёнными участками неба, получаются через сервисы VizieR и MAST и передаются в локальные центры обработки данных, где применяется сервис кросс-отождествления источников излучения в различных диапазонах (например, сервис SAI CAS, Россия). Далее необходимо оценить степень переменности звёзд по множественным наблюдениям в одних и тех же диапазонах в разные эпохи, а также на основании присутствия звёзд в базах данных переменных звёзд ОКПЗ, AAVSO (Россия, США) и др. Эти данные также перемещаются в соответствующие локальные центры и отождествляются с присутствующими фотометрическими данными. Таким образом, собираются фотометрические данные (до 15 полос) о блеске отождествлённых звёзд. Затем для параметризации звёзд используются результаты моделирования эволюции звёзд, рассчитанные программными комплексами PARSEC, COLIBRI, Y2 (Европа, США), и звездных атмосфер, например, рассчитанных комплексом ATLAS9 (США). В частности, данные расчётов собираются в SVO Theoretical Services (Испания) и могут пополняться за счёт библиографического поиска ADS и arXiv. Сервис, реализующий методику параметризации звёзд (ИНАСАН, Россия), использует собранные данные фотометрии для всего множества звёзд и данные моделирования спектров и оценивает параметры звёзд (включая расстояние до них) и межзвёздной среды. Результаты работы сервиса параметризации поступают в репозитории с каталогом астрофизических параметров звёзд и трёхмерной картой межзвездного поглощения. Точность работы сервиса параметризации звёзд по многополосной фотометрии оценивается сравнением с данными спектроскопических обзоров SEGUE, LEGUE (США, Китай) и картами межзвёздного поглощения, включая разработанную для европейского космического эксперимента

Hipparcos (Arenou et al, 1991). Таким образом, результатом работы комплекса станут каталог звёзд с астрофизическими характеристиками и расстоянием до них, списки объектов, не поддавшихся параметризации (вероятные кандидаты в новые классы объектов), и карта зависимости межзвездного поглощения от расстояния в различных направлениях. Эти данные используются для решения множества других задач, в частности, для уточнения параметров ускоренного расширения Вселенной. Данные карты межзвёздного поглощения в направлениях известных сверхновых перемещаются в центр обработки данных. На основании этих данных производится усреднённая минимальная оценка поглощения средой до сверхновых, а это позволяет уточнить плотность тёмной энергии, ответственной за ускоренное расширение Вселенной.

Неорганическая химия и материаловедение

НХМ 1. Термоэлектрические материалы для прямого преобразования тепла в электричество

В следующем десятилетии в связи с уменьшением запасов углеводородов и роста загрязнения окружающей среды особенно остро станет проблема «чистой» энергетики. Один из путей решения проблемы – использование термоэлектрических генераторов энергии, предназначенных для прямого преобразования тепла в электричество (Snyder G. J. Thermoelectric Energy Harvesting, in "*Energy Harvesting Technologies".* Ed. by S.Priya. Springer. 2009. P.325-336). Основным активным компонентом таких генераторов являются термоэлектрические материалы (ТЭМ).

Для выбора ТЭМ с широким набором заданных свойств (высокими коэффициентом термо-ЭДС и КПД, широким интервалом рабочих температур, низкой стоимостью, технологичностью в производстве, малой токсичностью, хорошими механическими свойствами и т.д.) необходимо провести поиск информации во множестве баз экспериментальных данных:

- термоэлектрические свойства – БД NIST ITS-90 Thermocouple Database (https://srdata.nist.gov/its90/main/), SpringerMaterials (http://materials.springer.com/) и БД, разработанная в Калифорнийском (Санта-Барбара) и Гарвардском университетах;
- токсичность - БД Chemical Safety Documents (SpringerMaterials);
- механические свойства – БД SpringerMaterials

и электронных коллекциях публикаций (ЭКП) (Elsevier (http://www.sciencedirect.com), Wiley (http://www.interscience.wiley.com), Springer (http://www.springerlink.com/), ACS (http://pubs.acs.org), Taylor&Francis (http://www.tandfonline.com) и т.д.). В случае отсутствия экспериментальных данных будет проводиться поиск результатов в базах расчетных данных (например, в Materials Encyclopedia, разрабатываемой в рамках NoMaD (http://nomad-lab.eu/), CompES (http://compes-x.nims.go.jp/index_en.html) и т.д.). В случае отсутствия последних, если это возможно, будет проведен расчет необходимых свойств с использованием квантовомеханических пакетов, например, VASP (https://www.vasp.at), CASTEP (http://www.castep.org/) и т.д.), а также методов data mining (методы распознавания образов по прецедентам – метод случайного леса, использованный в [Gaultois M. W., Oliynyk A. O., Mar A., et al. Web-based machine learning models for real-time screening of thermoelectric materials properties // APL Materials. 2016. V.4. N.5. P.053213], комплекс программ распознавания, включенный в систему [Kiselyova N. N., Dudarev V. A., Stolyarenko A. V. Integrated system of databases on the properties of inorganic substances and materials // High Temperature. 2016. V. 54. № 2. P.215-222], и т.д. Данные расчетов в дальнейшем помещаются в соответствующие БД с расчетной информацией.

Близкая по назначению задача, целью которой является энергосбережение, решается при использовании широкозонных полупроводников (ШП), которые позволили заменить обычные лампы накаливания на светодиодные источники света, что стало одним из прорывных достижений в развитии энергосберегающих технологий. Основными характеристиками ШП является ширина запрещенной зоны, интервал рабочих температур (БД Bandgap, SpringerMaterials, NSM (www.ioffe.ru/SVA/NSM/), CINDAS (https://cindasdata.com/)), подвижность носителей тока (БД NSM), температура плавления (БД NSM, Фазы и Диаграммы в (Kiselyova N. N., Dudarev V. A., Stolyarenko A. V. Integrated system of databases on the properties of inorganic substances and materials // High Temperature. 2016. V. 54. № 2. P.215-222), AtomWork (http://crystdb.nims.go.jp/index_en.html), Термические Константы Веществ (ТКВ) (http://www.chem.msu.su/cgi-bin/tkv.pl?show=welcome.html/welcome.html) и др.), токсичность, механические свойства (БД о последних двух свойствах, ЭКП, БД с расчетной информацией, в которых следует проводить поиск расчетных данных, а также список инструментария для проведения расчетов приведен выше).

НХМ 2. Кристаллы с особыми пьезоэлектрическими, электрооптическими и нелинейнооптическими свойствами

Кристаллы с особыми пьезоэлектрическими (ПЭ), электрооптическими (ЭО) и нелинейнооптическими (НО) свойствами широко используются в современной оптоэлектронике и лазерной технике в качестве рабочих тел оптических модуляторов, затворов, фильтров, приборов УЗИ, умножителей частоты и других преобразователей параметров светового пучка (в особенности лазерного излучения) и т.д. Особенностью этих кристаллов является отсутствие центра симметрии. Следует отметить, что это условие не является достаточным для проявления ПЭ, ЭО и НО эффектов, однако поиск и прогнозирование еще не полученных неорганических веществ с ацентричной кристаллической решеткой позволяют резко сократить затраты на разработку новых ПЭ, ЭО и НО материалов.

Для поиска экспериментальной информации об ацентричных кристаллах обычно используют базы кристаллографических данных: в большинстве случаев ISCD (http://www.nist.gov/srd/nist84.cfm), а также SpringerMaterials, AtomWork, Фазы, Кристалл и т.д. Если нужная экспериментальная информация не найдена, то применение методов распознавания образов по прецедентам позволяет успешно прогнозировать еще не полученные соединения с ацентричной кристаллической структурой.

Климатология

К1. Обработка и анализ многомерных пространственных данных по климату

Для климатической тематики Европейским центром среднесрочных прогнозов погоды ECMWF создаются веб сервисы для доступа к климатическим данным, их использования, анализа и визуализации (http://www.earthserver.eu/services/csds). Помимо этого, ECMWF отвечает за развитие климатических сервисов в климатическом блоке Copernicus climate change service (C3S, https://climate.copernicus.eu/about-c3s) Европейской программы спутникового зондирования Земли Copernicus (http://www.copernicus.eu/). Сервис C3S будет комбинировать спутниковые наблюдения и результаты моделирования для получения детальной информации о пространственных характеристиках климата прошлого, настоящего и будущего. Такой подход даст полное согласованное и надежное описание климата и тех его характеристик, которые важны для разработки мер по адаптации секторов экономики к происходящим климатическим изменениям.

Цепочку возможного использования такого ресурса можно описать следующим образом. Для весенней оценки возможного урожая с конкретного поля с помощью тематических веб-

сервисов готовится климатический анализ нормы и экстремалей для этой территории. При необходимости запускается метеорологическая модель и выполняется масштабирование метеохарактеристик реанализа ERA5 http://climate.copernicus.eu/climate-reanalysis с пространственным разрешением 31 км до необходимого пространственного разрешения. Запускается вычислительная модель урожайности конкретной культуры, в которой используются средние и экстремальные метеохарактеристики. Получаемые в результате оценки ожидаемой урожайности для различных культур дадут возможность принять обоснованное решение об оптимальном использовании конкретного поля.

К2. Изучение конкретной реакции наземных экосистем на различные типы климатических экстремальных явлений

Веб-ГИС «Климат» (http://climate.climate.scert.ru/) обеспечивает поддержку междисциплинарных исследований изменений регионального климата и отклика окружающей среды на них. В частности, для изучения конкретной реакции наземных экосистем на различные типы климатических экстремальных явлений волны жары, холодные периоды, сильные дожди или снегопады, штормы, наводнения или засухи), которые оказывают сильное воздействие на наземные экосистемы. Чтобы продемонстрировать, как лицо, принимающее решение, может использовать эту систему, предлагается следующий сценарий в области наук о Земле, который, по нашему мнению, может быть реализован в 2020 г. Сначала выполняется детальное изучение пространственно-временной динамики недавних климатических экстремальных явлений на территории Северной Евразии на основе доступных данных метеорологических (данные ВНИИГМИ-МЦД, http://meteo.ru/) и спутниковых (ESA Sentinel, https://sentinels.copernicus.eu/) наблюдений и результатов реанализов (ECMWF ERA5, https://climate.copernicus.eu/climate-reanalysis#table) и климатического моделирования (CMIP6) Анализируются статистические и динамические аспекты различных климатических характеристик и строится статистическая модель, описывающая взаимоотношения между климатическими экстремальными явлениями и экстремальными откликами экосистем на территории Северной Евразии. Результаты исследования динамики экстремальных явлений, полученные с помощью данных наблюдений, сравниваются с результатами, полученными с помощью региональных климатических моделей эксперимента CORDEX (http://www.cordex.org/), моделями WRF (http://www.wrf-model.org/) и ИВМ РАН СМ6, а также в проекте CMIP6 (https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6). Полученные результаты используются для создания входных данных для моделей динамики растительности (LPJmL и JSBACH). Полученные поля растительности детально оцениваются с точки зрения прогнозируемых экстремальных климатических и экосистемных явлений. В результате получается набор карт экосистемных рисков и уязвимостей для различных типов экстремальных климатических явлений на территории Северной Евразии (включая оценки периодов возврата для определенных бедствий и оценки неопределенности для различных типов экстремальных явлений) и возможный в будущем ущерб от них. Эти карты используются лицами, принимающими решения для управления природными ресурсами и сельским хозяйством. Карты представляются на веб-сайте в виде слоя для ГИС, что облегчает их использование заинтересованными потребителями.

11. Заключение

Анализ развития в мире е-инфраструктур обеспечения совместного доступа к данным в распределенной мультидисциплинарной среде исследовательских инфраструктур как необходимых средств поддержки Открытой науки показал следующее.

Евросоюз, следуя политике интеграции государств Европы, предпринимает естественные для подобной политики активные и целенаправленные действия по созданию и развитию средств

поддержки Открытой науки, преодолевая межгосударственные границы и барьеры. Согласно проведенному анализу, опыт таких консорциумов как, например, EGI, EUDAT, заявки на новые проекты образуют нужные предпосылки для создания Европейского облака открытой науки (EOSC). Прошедшие годы активного обсуждения проекта EOSC показали, что накопленный технический опыт достаточен для того, чтобы принять решение и запустить проект с тем, чтобы в 2020 году можно было привести в действие первую реализацию EOSC. Поэтому объявление Еврокомиссией в конце Октября 2017 г. программы реализации EOSC представляется своевременным и обоснованным.

Следует заметить, что накопленный опыт создания и использования инфраструктур данных и облачных архитектур в различных научных проектах Европы позволит распространить этот опыт на сферу экономики, основанной на данных и знаниях, что составляет важную проблему для развития Евросоюза. Иными словами, создание EOSC важно не только для науки, но и для экономики в целом.

Вместе с тем, видны неизведанные еще проблемы. Прежде всего, это вопросы семантики данных в мультидисциплинарной среде. Представляется, что решение следовать принципам FAIR правильное, однако, пока опыт реализации таких принципов на базе рекомендаций Семантического Веба, и, главное, опыт широкомасштабного внедрения такой реализации в практику (очень непростой для использования специалистами различных ОИИД) отсутствует. В ряде инфраструктур Европейской науки начаты исследования по расширению EUDAT поддержкой принципов FAIR для обеспечения семантической интероперабельности частей решения междисциплинарных задач. Первые результаты ожидаются к 2018 году.

Проект NDS, реализуемый в США, с точки зрения требований к совместному использованию разнообразных мультидисициплинарных данных близок к EOSC. Однако, пока не ясно, будут ли в проекте радикально решены проблемы семантики данных. Насколько известно, имеющийся в США практический опыт согласования семантики мультидисициплинарных данных ограничен средствами системы CINERGI проекта EarthCube. В целом, проект NDS развивается прагматически быстро, так что можно ожидать первых результатов до 2020 года.

Важную часть обзора составляют примеры (use cases) постановок задач в областях астрономии, материаловедения, климатологии, требующих совместного использования данных, имеющихся, а также ожидаемых после 2020 г. в мире и в России. Эти примеры нужны для взаимодействия с Европейскими проектами (прежде всего, с EOSC). Например, очевидным представляется предложение по исследованию возможности спецификации этих use cases согласно принципам FAIR для оценки идей реализации таких принципов, вырабатываемых для обеспечения семантической интероперабельности в различных инфраструктурах данных.

Программа создания EOSC, объявленная Еврокомиссией, потребует серьезных усилий сообществ различных мультидисциплинарных специалистов Открытой Науки в Европе. Судя по замыслу EOSC, успех этой программы позволил бы в основном решить технические проблемы обозначенных в [1] глобальных тенденций создания массивных коллекций данных в мире и обеспечения возможности совместного использования таких коллекций при решении задач исследования и принятия решений в различных ОИИД в России. Вместе с тем, большое число трудно решаемых организационных и правовых вопросов останутся открытыми.

Литература

1 Л.А. Калиниченко, Е.П. Гордов, Н.Н. Киселева, О.Ю. Малков, С. Позаненко, и др.(всего 12 авторов). Проблемы доступа к данным в исследованиях с интенсивным использованием данных в России, Информатика и ее применения, 2016, т. 10, Вып. 1, с. 3-23.

2 Guidelines on FAIR Data Management in Horizon 2020. European Commission, Version 3.0, 26 July, 2016.

3 Open Research Data in HORIZON 2020, http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf#view=fit&pagemode=none)

4 ESFRI 2016 Road Map (https://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/esfri_roadmap_2016_adopted.pdf)

5 HORIZON 2020-Work Programme 2016-2017. European Research Infrastructures (including e-Infrastrucures) http://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-infrastructures_en.pdf

6 HORIZON 2020-Work Programme 2016-2017. EINFRA 12-2017 и EINFRA 21 – 2017. http://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-infrastructures_en.pdf

7 European Open Science Cloud, https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

8 European Cloud Initiative, https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

9 Open Science, https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

10 EGI. - https://www.egi.eu/about/

11 EUDAT. - http://www.eudat.eu

12 EUDAT2020 – March 2015 – Feb 2018. - https://www.eudat.eu/eudat2020-%E2%80%93-march-2015-%E2%80%93-feb-2018

13 West-Life project. - http://about.west-life.eu/

14 HORIZON 2020-EINFRA 2016-2017 Call. - https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/calls/h2020-einfra-2016-2017.html#

15 Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

16 Force 11: GUIDING PRINCIPLES FOR FINDABLE, ACCESSIBLE, INTEROPERABLE AND RE-USABLE DATA PUBLISHING VERSION B1.0. - https://www.force11.org/fairprinciples

17 Starr J. et al. Achieving human and machine accessibility of cited data in scholarly publications // PeerJ Computer Science. – 2015. – Т. 1.

18 Mark D. Wilkinson, et al. Interoperability and FAIRness through a novel combination of Web technologies, October 13, 2016, https://peerj.com/preprints/2522v1

19 European Cloud Initiative – Building a competitive data and knowledge economy in Europe, European Commission, Brussels, April 19, 2016

20 GO-FAIR Application doc, June 2016, http://www.dtls.nl/wp-content/uploads/2016/06/Aanmeldformulier-Nationaal-Icoon-DTLS_signed.pdf

21 Realising the European Open Science Cloud, Commission High Level Expert Group on EOSC, Brussels, October 2016. -

http://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf#view=fit&pagemode=none
22 An EUDAT-based FAIR Data Approach for Data Interoperability, https://www.eudat.eu/communities/an-eudat-based-fair-data-approach-for-data-interoperability
23 NDS charter. - http://www.nationaldataservice.org/docs/Charter_v2.pdf
24 NDS Technology Components. - https://nationaldataservice.atlassian.net/wiki/display/NDSC/Technology+Components
25 Metadata Standards Directory, RDA. - http://rd-alliance.github.io/metadata-directory/
26 Core Scientific Metadata Model (CSMD). - http://icatproject-contrib.github.io/CSMD/
27 Н. А.Скворцов, Е. А.Аввакумова, Д. О.Брюхов, А. Е.Вовченко, А. А.Вольнова, О.Б. Длужневская, П. В. Кайгородов, Л. А. Калиниченко, А. Ю. Князев, Д. А. Ковалева, О. Ю. Малков, А. С. Позаненко, С. А. Ступников. Концептуальный подход к решению задач в астрономии // Астрофизический Бюллетень – Т 71. – 2016. – С. 122-133.
28 Toshihiro Ashino. Materials ontology: an infrastructure for exchanging materials information and knowledge // Data Science Journal. – 2010. – Т. 9. – С. 54-61.
29 P. Jiang, M. Elag, P. Kumar. GeoSemantic Resource Alignment Service, // CSDMS 2015.